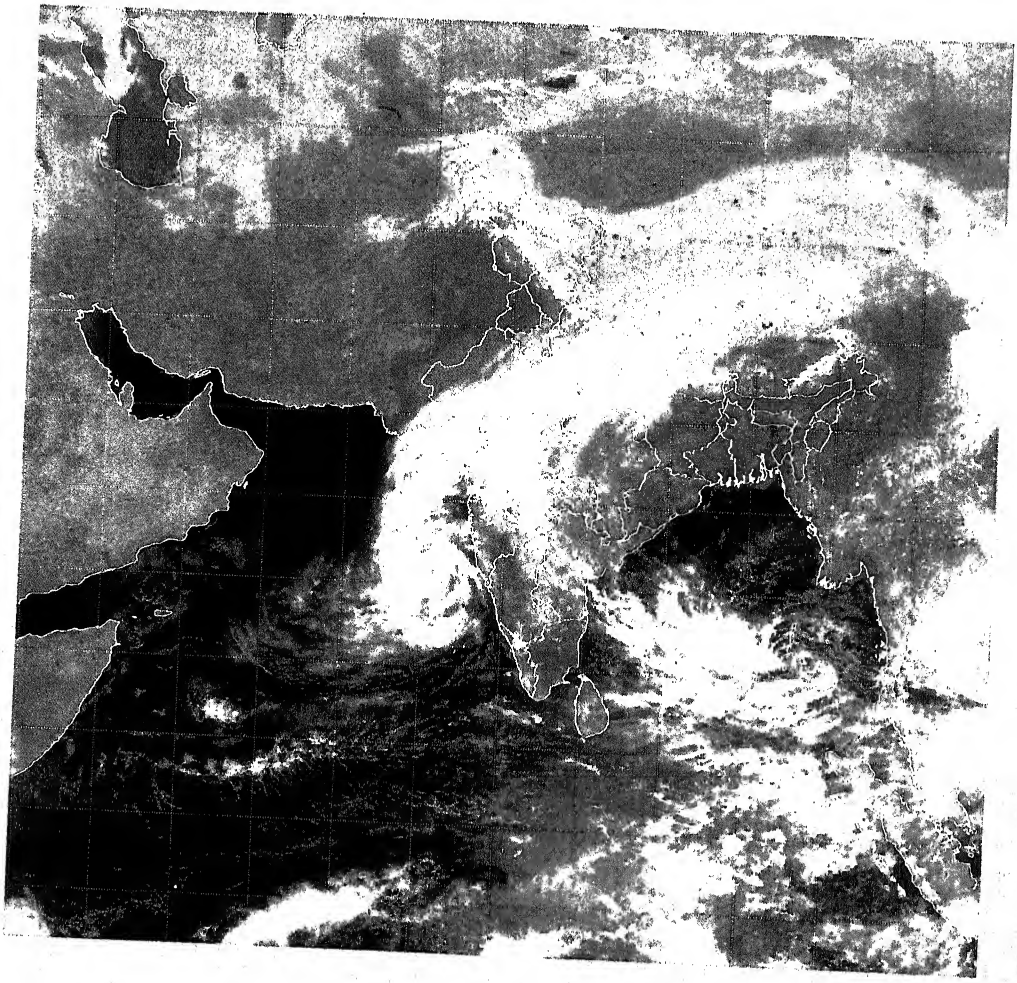
Resonance

November 1996

Volume 1 Number 11

journal of science education



The Challenge of Weather Prediction ❖ Combinatorial Group Theory❖ Flash Memories ❖ Cryptatium: The First Elementoid ❖ Talipot and its Conservation

Resonance is a monthly journal of science education published by Indian Academy of Sciences, Bangalore, India.

Editorial Office: Indian Academy of Sciences, C V Raman Avenue, PB No. 8005, Bangalore 560 080, India.
Tel:+91 (80) 3342310 / 3342546, Fax:+91 (80) 3346094, email:resonanc@ias.ernet.in

Editorial

N Mukunda, Chief Editor

It is generally true, particularly in our country, that students and teachers of science, and even working scientists, pay little attention to matters of the history and philosophy of science. Of course some degree of maturity is required before one sees the need for and value in studying these subjects ; and in most of our institutions they are yet to take root as serious pursuits. In this issue we feature Karl Popper's portrait on the back cover, and two brief *articles in boxes* on Popper by Gangan Prathap and M G Narasimhan. Popper's name has appeared in some earlier *Resonance* articles – notably Bondi's "Science-its Philosophy and Spirit " in July 1996 – and he is widely regarded as about the most influential philosopher of science of the century. Prathap's article explains why. At the same time we would like to impress upon our younger readers that *Popperian ideas* – howsoever influential – are part of an evolving discipline, and should be approached with a critical and open mind. It is to reinforce this that we requested M G Narasimhan to tell us briefly why life scientists have been quite critical of Popper's formulations. In spite of Sellars and Yeatman, history is never a finished subject, it is always a recreation of the past influenced by new findings and new viewpoints. So is it with the history and philosophy of science as well.

History is never finished, it is always a recreation of the past influenced by new findings and viewpoints.

Turning to other pieces – Goswami begins to explain the difficulties of weather prediction. This reminds us of Niels Bohr's dictum – "It is difficult to predict, especially about the future". Nagesh Rao brings out the structure behind the seemingly simple procedures of dimensional analysis; and Amitabh Joshi contrasts the older *gerontological* and the newer *evolutionary* approaches to the problem of ageing. So, read on and stay young!

"It is difficult to predict, especially about the future"
– Niels Bohr

Karl Popper

Karl Popper was arguably the greatest philosopher of science of the century; some say, the greatest philosopher of the century. Although he is not the household name he deserves to be, he is not entirely unfamiliar to the *Resonance* readership. He has been frequently cited in the pages of *Resonance*, in Narlikar's series on The Origin of the Universe, in Bondi's statement on the spirit of science, etc.

Sir Karl Raimund Popper was born on 28 July 1902 in the Ober St Velt district of Vienna. His father was a scholarly and distinguished lawyer with radical liberal sympathies. His mother was a talented musician and from her he inherited a love for music. After school, he studied mathematics, physics, psychology, philosophy and music, earning his doctorate in 1928 from the University of Vienna. After qualifying as a secondary school teacher, he taught at a high school until 1937.

During this period, he actively pursued his studies on the nature and theory of knowledge, and his first book, *Logik der Forschung*, was to contain his solutions to some of the fundamental problems of the Theory of Knowledge. The success of this book established Popper's reputation as a philosopher. This book laid down the foundations for his unending crusade for critical rationalism, for unraveling the nature of the scientific enterprise and for defining the scientific method.

So, what is the scientific method, according to Popper? The traditional view, which goes back to Bacon, Galileo and Newton is that observations and experiments come first. By induction, general theories are ground out from these observations and hypotheses are then derived from these. Further experiments are performed to verify these hypotheses. The original theory is thus claimed to have been proved or disproved and the scientist assumes she has arrived at the truth of the matter until proven wrong.

Popper did not like the idea of experiment as a proof of a theory. He believed instead that wherever there was a recognisable problem (the theory did not fit in with the facts), a bold and imaginative theorist will make a risky conjecture (the riskier, the better) and then try hard to design experiments to falsify predictions made from this conjecture; i.e. carry out tests or experiments not to prove the predictions but actually to *refute* them. The refutations will help to modify the original conjecture and lead to a better approximation.

Empirical tests are therefore used to weed out falsehoods (not prove truths!). In fact, Popper believed, the truth could never be conclusively achieved; only closer and closer approximations to the truth are realised. A single counter-example can show a *law* to be false.

The role of experiments in all this is very clear. Popper has argued that "in the history of science, it is always the theory and not the experiment, always the idea and not the observation, which opens the way to new knowledge . . . it is always the experiment which saves us from following the track that leads to nowhere; which helps us out of the rut, and which challenges us to find a new way."

The violent turn of political and social events in Europe triggered off by Hitler's Nazism in Germany forced Popper to move first to Cambridge and then to New Zealand in 1937. He remained there till 1945, teaching philosophy. While there, he wrote a famous argument for democracy, moderation and liberality which also served as a ferocious attack on totalitarianism, titled *The Open Society and its Enemies*. It is a covenant of principles by which a community can ensure its own well being and that of its children.

Popper believed that it is individual human actions, with consequences which are more often unintended than intended, which produce the spontaneous order that we recognise as civilised society. The central point of good rule is not "Who should rule?" but "How can we get rid of bad rulers without bloodshed?" Popper's liberalism is based on utility; his articulation of the principle of utility is governed by the negative formulation, *eliminate suffering* and not *maximise happiness*. His views on social reform thus mirror his views on nature and science.

He returned to England in 1946, to teach logic and scientific method at the London School of Economics and Political Science till his retirement in 1969. He was decidedly the most important mind to have worked at the School. He was elected a Fellow of the Royal Society in 1976. He passed away, at the age of ninety-two, on 17 September 1994.

Ray Percival in a review in *New Scientist* said that "Apart from Aristotle and Plato, no other thinker can equal the breadth and depth of Popper's contributions to knowledge. Popper's mind grappled with problems ranging from logic and quantum physics to evolutionary, social and political theory. And in a century obsessed with specialisation, breadth is astonishing."

Let me end this piece with a voice from our own country. T G Vaidyanathan wrote on Popper in the pages of *The Hindu* in October 1994: "We should make the study of Popper compulsory in our degenerating universities to restore rationality and universalism so essential to the creation of an open society anywhere and everywhere. The alternative roads, touted from every rabble-rousing platform in India, all lead to serfdom."

G Prathap

Popper on Darwinism

In his philosophical analysis of the nature of scientific knowledge Popper differentiated between scientific theories and metaphysical theories or research programmes in terms of their testability. Popper felt that only such theories could be legitimately termed scientific which could be tested and refuted by experience.

Known as the *Demarcation Criterion,* this differentiating principle rendered Darwin's theory of Organic Evolution by natural selection metaphysical as it did not lend itself to such refutation. In Popper's view, the utmost that could be said about Darwinism (this includes neo - Darwinism) is that while in itself it was not testable, it offered a possible frame work for testable scientific theories. The crux of Popper's argument lay in the observation that the evolutionary theory was incapable of making any prediction either in principle or in practice. As a result, the theory had limited explanatory ability in comparison with other scientific theories such as Newton's theory of gravitation. In sum, then, for Popper Darwinism was a metaphysical theory or a research programme, although a very special kind of metaphysical theory because it had provided valuable explanations in understanding at least certain aspects of natural phenomena.

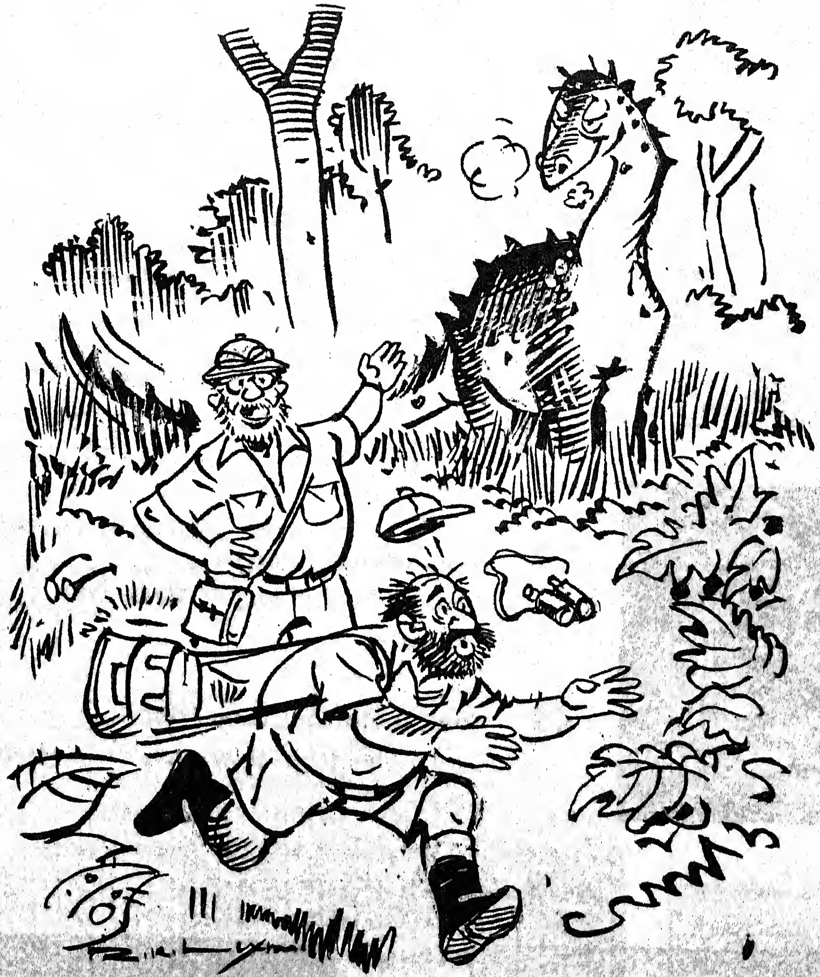
Popper's description of Darwinism as a metaphysical research programme was subjected to severe criticism by both biologists and philosophers of biology. Later, Popper modified his position by admitting that the theory of evolution could be often tested by deriving testable predictions from it. But philosophers of biology like Michael Ruse have argued that this modification did not amount to any substantive change in Popper's overall position. Ruse in particular argued that Darwinism taken as a whole was capable of making predictions as in the case of *founder principle,* an important principle in Population Genetics. He also argued that in terms of Darwinism one could claim that a species once having become extinct is unlikely to come back. In other words, according to Darwinism evolution cannot repeat itself. As Ruse puts it in his book *Darwinism Defended* (published by Addision Wesley, 1983) " the dodo is gone for ever".

Notwithstanding these limitations of Popper's views on Darwinism, his general philosophy of science does incorporate certain important concepts from the theory of evolution . For instance, Popper's model of scientific change emphasizes the *natural selection* of a given hypothesis in the context of competing hypotheses and the gradual replacement of one hypothesis by another on the basis of empirical refutation. This model of scientific change based on what Popper calls *Conjectures and Refutations* has been taken as the foundation of a new discipline in philosophy known as *Evolutionary Epistemology.*

M G Narasimhan

Science Smiles

R K Laxman



Don't be silly, Professor! All our research and study have shown this was extinct some two million years ago!

Resonance journal of science education

November 1996 Volume 1 Number 11

SERIES ARTICLES

8 **The Challenge of Weather Prediction**
The Basic Driving
B N Goswami

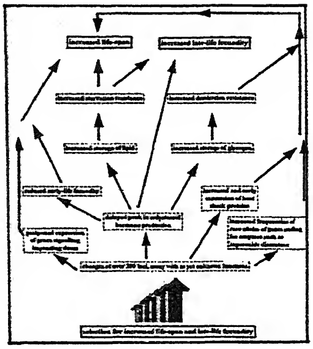
18 **Know Your Personal Computer**
The CPU Base Instruction Set and Assembly Language Programming
S K Ghoshal

GENERAL ARTICLES

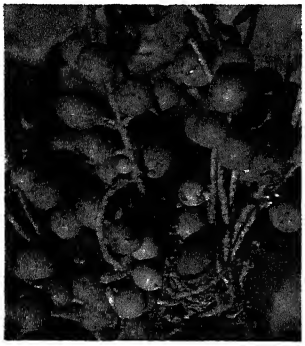
29 **Dimensional Analysis**
Keeping Track of Length, Mass, Time ....
N N Rao

42 **Combinatorial Group Theory**
Group Theory via Generators and Relations
B Sury

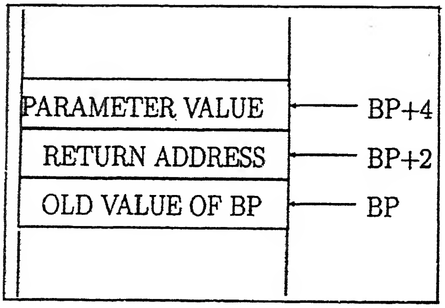
51



69



18



64

51 **Evolution, Fruit Flies and Gerontology**
Evolutionary Biology Helps Unravel the Mysteries of Ageing
Amitabh Joshi

64 **Aerogel**
The Lightest Solid Known
D Haranath

FEATURE ARTICLES

69 **Talipot: A Forgotten Palm of the Western Ghats**
A Plea for its Conservation
M D Subash Chandran

76 **What's New in Computers**
Flash Memories
Vijnan Shastri

83 **Molecule of the Month**
Cryptatium, the first "Elementoid"
Uday Maitra

DEPARTMENTS

Editorial 1
Chief Editor's column / Karl Popper... *G Prathap and M G Narasimhan* / Science Smiles *R K Laxman*

Classroom 88
Bachet's Problem
B Bagchi

Think It Over 92
Questions raised in Vol.1, Nos. 6 & 7.

Information and Announcements 93
National Seminar for Science Writers/ 1996 Nobel Prizes.

Books Received 94

Front Cover
An INSAT ID cloud picture of the October 1996 Arabian Sea Cyclone taken at 0600 GMT on October 23, 1996. It covers the region between 10° S and 45° N and 44° E and 105°E. Another weaker vortex is noted around Andaman-Nicobar islands. These vortices are embedded on the large scale Inter Tropical Convergence Zone (ITCZ) from Indonesia to Arabian Sea. (Courtesy India Meteorological Department)

Back Cover
Karl Raimund Popper (1902-1994)
(illustration by Prema)

The Challenge of Weather Prediction

1. The Basic Driving

B N Goswami

B N Goswami is with the Centre for Atmospheric and Oceanic Sciences at the Indian Institute of Science, Bangalore. After his PhD in Plasma Physics he was attracted to this field by the challenges in weather and climate prediction. His interests include understanding variability and predictability of all tropical phenomena including the monsoon.

This three part series brings into focus problems and challenges involved in weather forecasting. The first part deals with the fundamental forces governing weather and climate while the second part deals with the practical and conceptual difficulties. The last part describes old and new ways of weather forecasting.

Happy with the forecast from the meteorological department for a sunny day, you decided to take your family to the beach only to be drenched by a heavy afternoon thundershower! Such experiences have made weather forecasters a favourite subject matter for the cartoonist. But have we ever asked, why do the meteorologists go wrong? Is it because the meteorologists are incompetent or is it that the weather is intrinsically difficult to predict? Certain systems governed by physical laws can be predicted with precision well in advance, for example, the precise occurrence of the ocean tides or the positions of the planets around the sun. If weather is also governed by physical laws, why then is it so difficult to predict?

In this article, I want to enlighten you on certain common myths about weather forecasting and illustrate the inherent complexity of the problem. I will discuss what can be predicted and what cannot. I will also describe how weather forecasting evolved from an empirical science in the pre-second world war days to a quantitative analytical science in recent years.

The Basic Driving

To appreciate why it is so difficult to predict the weather, let us start by understanding what causes weather. The atmosphere is a gaseous (fluid) envelope around the earth. The weather is nothing but the day to day fluctuations of the atmospheric state. These fluctuations are due to the movement of air in the gaseous envelope. What causes the air to move?

The weather is nothing but the day to day fluctuations of the atmospheric state.

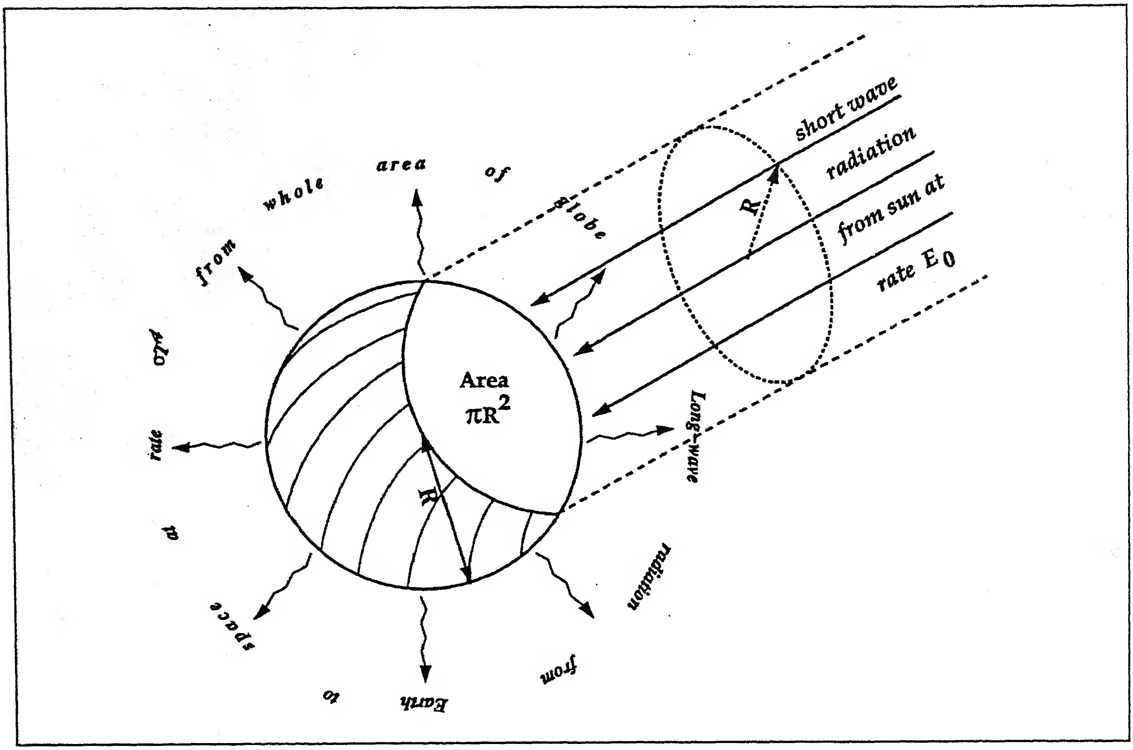
The air moves because it is under the action of a number of forces. The primary force acting on the atmosphere is solar heating which is an external force. As we know, hot bodies radiate according to Planck's law (higher the temperature of the black body, shorter the mean wavelength at which it radiates). As a result of the very high mean temperature of the sun, the solar electromagnetic radiation has maximum energy at the visible wavelengths. The gases in the atmosphere (mainly nitrogen, oxygen, water vapour, and carbon dioxide) cannot absorb much of this radiation. Part of it gets reflected from the clouds and the surface of the earth and the rest is absorbed by the solid earth. This heat raises the temperature of the earth's surface and it radiates as a black body. At the temperature of the earth-atmosphere system, the radiation emitted from the earth is concentrated in the infrared range. Measurements on earth indicate that the rate at which solar energy impinges on an area perpendicular to the sun's rays at the mean earth-sun distance is about 2.0 cal/cm^2/min or 1390 Watts/m^2. Let us call this the solar constant, S. Now, we can estimate the annual and global average equilibrium temperature of the earth (T_e) as a balance between incoming solar radiation and outgoing terrestrial (or emitted from the earth) radiation as shown in*Figure 1*.

The mean radiative equilibrium temperature of the earth can be estimated as a balance between incoming shortwave solar radiation and outgoing longwave radiation emitted by the earth.

The solar rays only intercept a disc of area πR^2 (where R is the mean radius of the earth) at any time. But the earth radiates in all directions with an area of $4\pi R^2$. A fraction α (called *albedo*) of the incident solar radiation is reflected back into space. This is found to be about 30% or 0.30. If the emissivity of the earth is ε and assuming that the atmosphere is also transparent to the longwave radiation, the balance demands that

$$4\pi R^2 \varepsilon \sigma T_e^4 = (1-\alpha)\pi R^2 S, \qquad (1)$$

Figure 1 Calculation of planetary temperature.



where α is the Stefan-Boltzmann constant(=5.6703 ×10^{-8}Wm^{-2} K^{-4}). Assuming that the earth is a perfect black body (ε= 1), the mean equilibrium temperature of the earth would beT_e =256°K (or −17°C). If this were the case, the whole earth would be ice covered and devoid of any life form! However, much to our delight the observed global average temperature is close to a comfortable 288°K (or 15°C). This is due to the *greenhouse effect* of the atmospheric gases. Some of the molecules of the atmosphere such as water vapour (H_2O) and carbon dioxide (CO_2) have natural oscillations with frequencies corresponding to the frequencies of the infrared waves emitted by the earth. When the electromagnetic waves (EM) fall on these molecules, they produce resonant excitation of these natural oscillations. In the process the EM waves lose energy. Thus, these gases are effective in absorbing the emitted infrared radiation. When the air absorbs such energy, it gets heated and radiates some energy back to earth and some to the outer space. Taking this into account, the infrared transmissivity Γ of the atmosphere is less than unity (in equation (1) Γ is 1). From the amount of H_2O and CO_2 in the atmosphere, it may be estimated that the infrared transmissivity of the atmosphere is $\Gamma \approx 0.62$. Therefore, the average surface temperature of the earth T_s may be estimated from the modified form of equation (1) as

$$4\pi R^2 \varepsilon \Gamma \sigma T_s^4 = \pi R^2 (1-\alpha) S, \qquad (2)$$

with $\varepsilon = 1$ and $\Gamma = 0.62$, $\alpha = 0.3$, the surface temperature isfound to be $T_s \approx 288$°K.

The earth's atmosphere is unique as it has just the right amount of greenhouse gas such as CO_2 and H_2O without which the equilibrium temperature would have been an inhospitable −17°C.

We note from equation (2) that there are two factors that can change the global average annual mean surface temperature T_s, the external solar forcing (S) and the infrared transmissivity (Γ) of the atmosphere. It is known that the incident solar energy changes slowly with periodicities of 22,000 years, 41,000 years and 100,000 years. These are due to changes in the earth's orbital parameters such as the eccentricity of the orbit (100,000 years), axial precession (22,000 years) and change in the obliquity of the ecliptic or the axial tilt (41,000 years). In fact paleoclimatic reconstruction shows fluctuations in the earth's temperature with such periodicities. Thus, if the constituents of the atmosphere never change (Γ constant), the prediction of the earth's long-term mean temperature would be easy (just like the tides)! However, the concentration of some of the constituent gases such as CO_2 is increasing in the atmosphere mostly due to human activity. As a result Γ is becoming smaller. From equation (2), it is clear that as Γ becomes smaller, the equilibrium surface temperature increases. This is the basis for the green house gas induced global warming. In fact, observed global mean temperature has shown an increase of a little over half a degree Celsius over the past hundred years.

If temperature was uniform everywhere on the earth, there would not be any motion. Although globally averaged temperature is a gross measure of the earth's climate, due to the

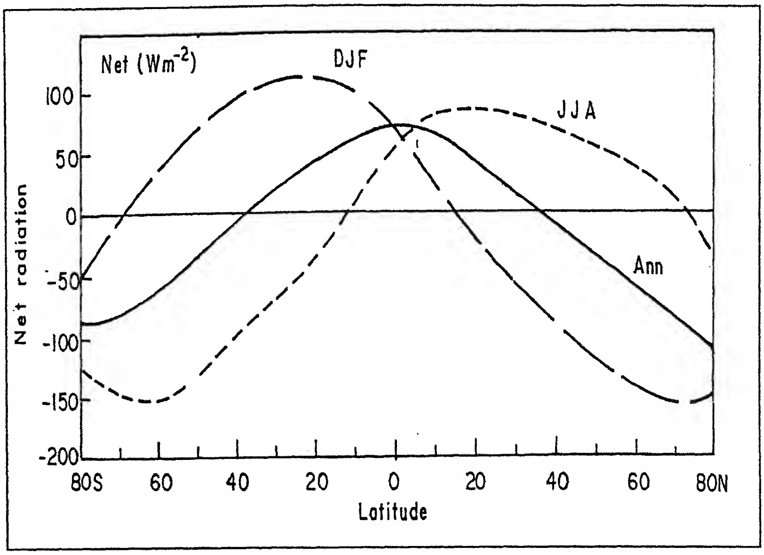


Figure 2 Net radiation (incoming shortwave-outgoing longwave) at the top of the atmosphere at different latitudes averaged over each latitude circle in the east-west direction. Annual mean and seasonal averages for June-July-August (JJA) and December-January-February (DJF) are shown. No corrections are made for global radiation balance.

geometry of the earth and its orbital revolution around the sun, the net radiation received at the top of the atmosphere is positive near the equator while it is negative near the poles (*Figure 2*) and varies with the season. This net radiative forcing is the result of a complex interaction of the incident solar radiation and the emitted longwave radiation with the atmosphere. *Figure 3* shows the processes involved in the shortwave radiation budget while*Figure 4* shows those associated with the longwave radiation budget. Each of these processes involves quite complex interactions between radiation and particles of different shapes and sizes in the atmosphere. For example, absorption of incident solar radiation by the atmosphere and clouds depends on the distribution of different gases and aerosols and droplet sizes in the cloud. Similarly reflection from the cloud also depends on the cloud characteristics. The reflection from the earth's surface depends on soil characteristics, distribution of vegetation, snow accumulation etc. Calculation of the longwave radiation budget is even more complex. For example, the absorption of longwave radiation by H_2O and CO_2 does not occur at one wavelength, but over a band of wavelengths. Therefore, one has to estimate how much is absorbed in all the wavelengths in the band and then sum them up. This is an involved process. In addition, a certain amount of energy is

Nonuniform radiative forcing (heating in the tropics and cooling in the polar regions) is the primary cause of atmospheric motions. Actual trajectory of an air parcel is determined by the balance between a number of forces such as the Coriolis force, the pressure gradient force, the gravitational force and the frictional force.

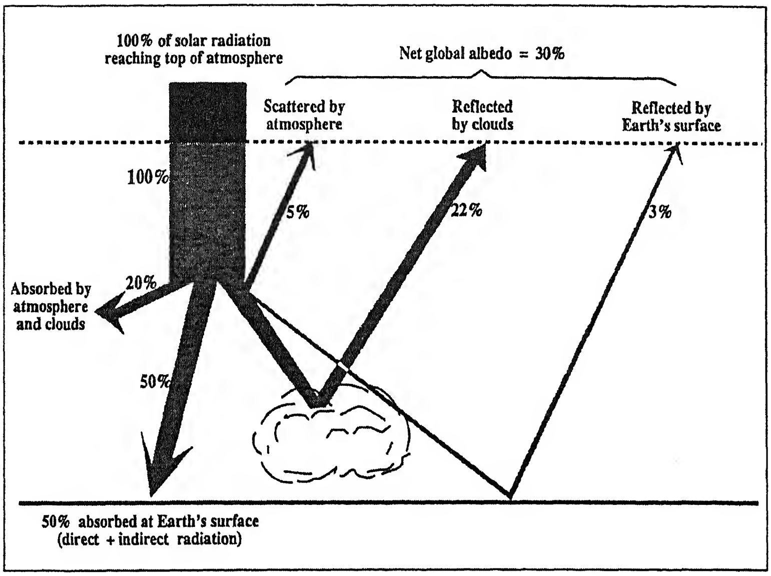
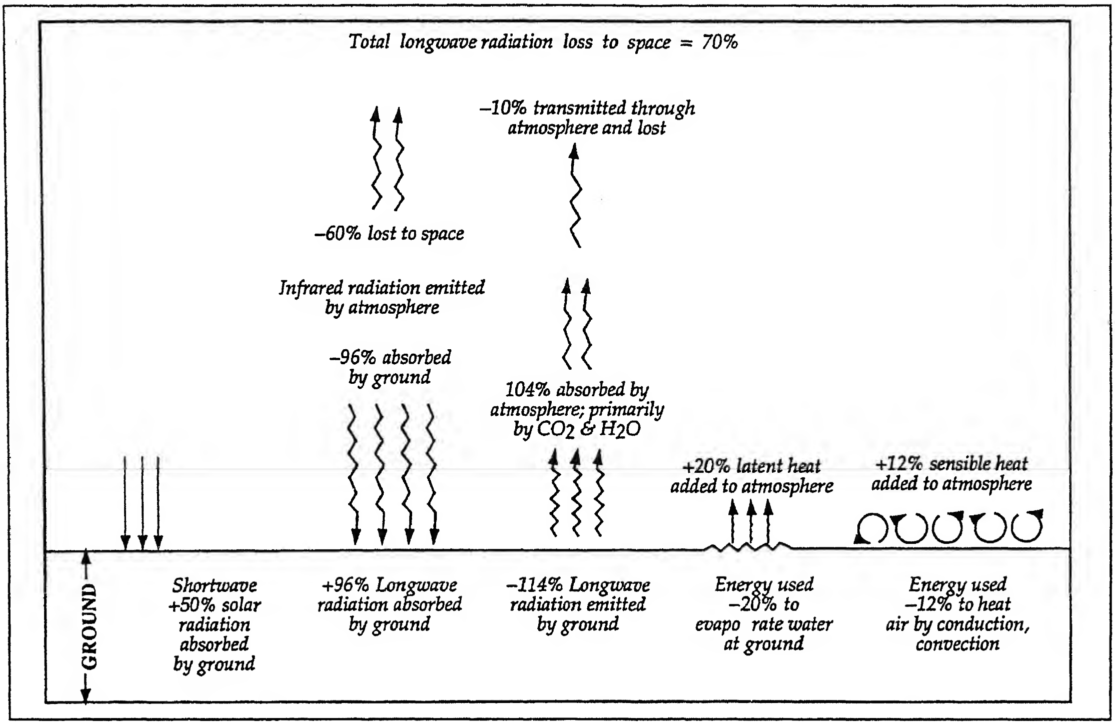

100% of solar radiation reaching top of atmosphere
Net global albedo = 30%
Scattered by atmosphere
Reflected by clouds
Reflected by Earth's surface
100%
5%
22%
3%
Absorbed by atmosphere and clouds
20%
50%
50% absorbed at Earth's surface (direct + indirect radiation)

Figure 3 Incoming solar radiation budget of the earth and the atmosphere.

exchanged between the atmosphere and the underlying surface. This exchange takes place through transport by conduction and convection in the atmosphere. However, as we shall explain shortly, it is not molecular diffusion but turbulent diffusion that achieves this. This exchange can be divided into two parts. One part is termed as *latent heat flux* while the other is called *sensible heat flux*. The word latent refers to energy stored in molecules that is released (or taken) when phase change takes place. For example, wind blowing against the surface of a wet land or the ocean, produces evaporation. During evaporation liquid water changes to water vapour. The land or the ocean loses the latent energy required for this change of state. The heat energy lost by the surface is gained by the atmosphere when the water vapour condenses back to water during the formation of clouds. The sensible heat flux refers to the exchange that takes place through conduction and convection when the atmosphere literally 'senses' the surface.

This nonuniform radiative forcing is the primary reason for motion in the atmosphere. Air in the hotter tropics becomes lighter and rises while the colder air in the polar region sinks and tries to flow to the tropics to replace the depleted air, thereby

Figure 4 Outgoing longwave radiation budget of the earth and the atmosphere.


Total longwave radiation loss to space = 70%
–10% transmitted through atmosphere and lost
–60% lost to space
Infrared radiation emitted by atmosphere
–96% absorbed by ground
104% absorbed by atmosphere; primarily by CO2 & H2O
+20% latent heat added to atmosphere
+12% sensible heat added to atmosphere
GROUND
Shortwave +50% solar radiation absorbed by ground
+96% Longwave radiation absorbed by ground
–114% Longwave radiation emitted by ground
Energy used –20% to evapo rate water at ground
Energy used –12% to heat air by conduction, convection

setting up a huge convection cell. Once the air is set into motion, it is affected by several other forces. They are:

- *Coriolis Force*: This force arises due to the rotation of the earth. Consider an air parcel moving towards the equator from the north pole. As the earth moves from the west to east and as the air is not rigidly attached to the surface, it will be deflected to its right. As a result, to an observer sitting on the surface of the earth it would appear to be blowing from northeast to southwest rather than from north. to south.

- *Pressure Gradient Force*: This force arises if the pressure at two points are different. Then fluid at high pressure tends to move to low pressure. When the tropics get heated, the air becomes light and pressure becomes low while the high latitude has a high pressure due to cold heavy air. So air would move from high latitude to low latitude.

- *Gravitational Force*: The air is also being continuously pulled by the gravitational force of the earth. This force acts vertically downward. One may ask, if the gravitational force is pulling the air all the time towards the centre of the earth, why then does all the air not accumulate near the surface? This is because, the gravitational force directed vertically downwards is normally balanced by a vertical pressure gradient force directed vertically upwards. We know that the air density decreases from the surface upwards. This sets up a pressure gradient force directed upwards from the surface. So, if the gravitational force were not there, the air from high pressure near the surface would tend to flow upwards towards the low pressure region. The gravitational force balances this tendency of the air and keeps the atmosphere in place. This balance may not hold always as in the case of a thunderstorm. Any imbalance between these two forces results in vertical motion.

- *Frictional Forces*: In addition to the above forces that tend to produce acceleration, there are forces that retard the motion. These are the frictional forces primarily active near the surface of the earth. The frictional forces essentially result in dissipation of energy. For a *laminar* flow above a rigid surface, this occurs through molecular diffusion. The molecules immediately in contact with the rigid surface will be slowed down. Then molecules just above them will be slowed down as they are in contact with these layers. This way, the solid surface would try to retard the flow up to a certain height. The rate at which the molecular diffusion slows down a flow is well known from laboratory experiments. Studies indicate that dissipation of energy and transfer of momentum near the earth's surface take place about a million times faster than molecular diffusion! How does this happen? This is because the typical atmospheric flow near the earth's surface is not *laminar* but *turbulent*. Such flow generates myriads of small scale (horizontal scale of the order of one meter) eddies. These eddies behave like big molecules and produce small scale vertical circulations that mix the air near the surface with the air above much more effectively than the

Due to turbulent eddies dissipation of atmospheric motion near the earth's surface takes place a million times faster than molecular diffusion.

individual air molecules. Therefore, in order to study how the exchange of momentum as well as heat takes place between the atmosphere and the surface below, we have to deal with the turbulent eddies!

Seven Equations and Seven Unknowns

Thus the evolution of the atmosphere from a given state to some future state is determined by certain laws of physics. For most, but not all, meteorological problems the relevant laws of physics can be expressed in terms of seven equations which govern the behaviour of seven variables (*Table 1*).

Table 1

Variables	Equations
1. Pressure (P)	Gas law or equation of state. Relates temperature, pressure and density, $P = \rho RT$.
2. Temperature (T)	First law of thermodynamics. Relates temperature changes to external heating or cooling, eddy diffusion and changes in pressure, external heating or cooling due to radiative balance and latent heat released during cloud formation.
3. Density (ρ)	Continuity equation for air; expresses conservation of mass of air.
4. Water vapour (q)	Continuity equation for water vapour; expresses rate of change of water vapour in terms of source (evaporation) and sink (precipitation) and diffusion due to turbulent eddies.
5. West to East component of wind (u)	Newton's second law (force = mass × acceleration) applied to west-east, south-north and vertical direction separately. The pressure gradient force, the Coriolis force, gravitational force and frictional force are taken into account.
6. South to North component of wind (v)	
7. Vertical component of wind (w)	

Suggested Reading

- Nigel Calder. *The Weather Machine.* The Viking Press. New York. p 144, 1974.
- John M Wallace and Peter Hobbs. *Atmospheric Sciences: An Introductory Survey.* Academic Press. San Diego, California. p 467, 1977.
- Richard A Anthes, John J Cahir, A B Fraser and Hans A Panofsky. *The Atmosphere.* (Third Edition) Charles E Merrill Publishing Company. Colombia, Toronto, London, Sydney. p 532, 1981.
- A Miller, J C Thompson, R E Petterson and D R Harayau. *Elements of Meteorology.* Charles E Merrill Publishing Company. Colombia, Toronto, London, Sydney. p 417, 1983.
- J P Peixoto and A H Oort. *Physics of Climate.* American Institute of Physics. p 520, 1992.

Address for Correspondence
B N Goswami
Centre for Atmospheric and
Oceanic Sciences
Indian Institute of Science
Bangalore 560 012, India
email:
goswamy@cas.iisc.ernet.in
Fax: (080) 334 1683

Know Your Personal Computer

5. The CPU Base Instruction Set and Assembly Language Programming

S K Ghoshal

Siddhartha Kumar Ghoshal works with whatever goes on inside parellel computers. That includes hardware, system software, algorithms and applications. From his early childhood he has designed and built electronic gadgets. One of the most recent ones is a sixteen processor parallel computer with IBM PC motherboards.

This article describes the instruction set of the base architecture by illustrating it with an assembly language program.

Instruction Set

The instructions supported in the 8088 can be classified into:

Data Transfer Instructions that move data from the source to the destination. Examples are MOV, PUSH, POP and XCHG.

Arithmetic Instructions like ADD, SUB, MUL, DIV, INC (which increments) and DEC (that decrements).

Logic Instructions like NOT, AND, OR, XOR, and TEST.

Bit manipulation Instructions like SHL, SHR, ROL and ROR.

String Instructions operate on string. MOVS copies strings, SCAS scans them, STOS initializes them and CMPS compares them. MOVS also provides an alternative to MOV instruction for moving data. Similar alternatives exist for other string instructions. This is typical of a CISC architecture. There is often more than one way to do the same thing.

Assembly Language Programming

We will illustrate with an assembly language program which implements a computable function that calls itself. It computes the factorial of a non-negative integer. The factorial function, as we know can be recursively defined as in equation 1.

$$f(x)=\begin{cases} 1 & \text{if } x=0 \\ x\,f(x-1) & \text{otherwise} \end{cases} \qquad (1)$$

Figure 1 An assembly language program to compute factorial.

```
.model small ; the virtual address space is small: only 64KB long
.code ; executable code follows in this assembly module
public _factorial      ;public near procedure callable by any C module
_factorial proc near ; subroutine entry point
push bp        ; save caller's BP
mov bp,sp ; fix the context of this instance
enter:              ; now the stack looks likeFigure 3
push bx        ; save old BX (that was the caller's copy of n)
push dx; save  DX too (MUL affects DX must restore to caller)
inside:              ; Now stack looks likeFigure 4 bottom-left
mov ax,[bp+4]  ; get parameter value from ss:[bp+4]
or ax,ax       ; check if the parameter is 0
jnz deeper     ; go deeper if parameter is non-zero
mov ax,1       ; 0! = 1
jmp done ; return one as result
deeper:           ; go deeper in recursion
mov bx,ax ; put parameter (n) in bx register
dec ax         ; calculate (n–1)
push ax        ; save (n–1)
call _factorial ; recursive call to itself
add sp,2       ; discard the parameter after return
mul bx         ; n * (n–1) –> dx and ax: dx gets msb, ax <– lsb
done  :              ; we are finished either way
pop dx         ; pop the saved registers dx first
pop bx         ; then bx – observe the reverse order
pop bp         ; restore caller's context
ret          ; return thread of control to caller
_factorial endp  ; the subroutine ends here
end ; the assembly module ends here
```

Figure 2 A C function to compute factorial.

```
int factorial (int arg1) /* An integer function with
                an integer argument */
{
  if (arg1= =0) return(1); /* If argument is zero,
                return 1 as result */
 return(arg1 * factorial(arg1–1)); /*Otherwise
                return arg1*factorial(arg1–1) */
}
```

Our implementation of the factorial function follows the same logic. This assembly language program (See *Figure 1*) is written in such a way that it can be called from C.

It is functionally equivalent to the implementation in C, as given in *Figure 2.* Note that comments (statements that are meant to be read by a human and not translated by a compiler are called comments) in C are between '/*' and '*/', whereas in assembly language whatever follows a semicolon in a line is a comment.

Case Sensitivity

If the use of upper-case or lower-case letters in a programming language changes the syntax and semantics of the program, then that language is called case-sensitive. C is case-sensitive. printf("Hello") is correct. pRintf("Hello") is wrong. Assembly language is case-insensitive. Both push bx and PUSH BX are correct and mean the same thing. Fortran 90 is not case-sensitive.

And as you have probably guessed, when you compile a program like the one in *Figure 2* (the exact text of your program will depend on the type of C programmer you are) it becomes a program like the one in *Figure 1* (the exact code produced will depend on the type of C compiler you use). There are variations in the details (for that matter the way one writes equation 1 depends on the person) of intermediate representation and the symbols used, but equation 1, *Figure 1* and *Figure 2* represent the same thing to different beholders. However, as the beholders understand different languages and operate at different levels, translation is required so that one beholder can work for the other. From C to assembly language, the translation process is called *compilation.* From assembly language, translation into machine language (some call it *binary* or *object code*) is by a program called the *assembler*. Only after that final translation step is done can the object code produced by the compiler be executed by the hardware.

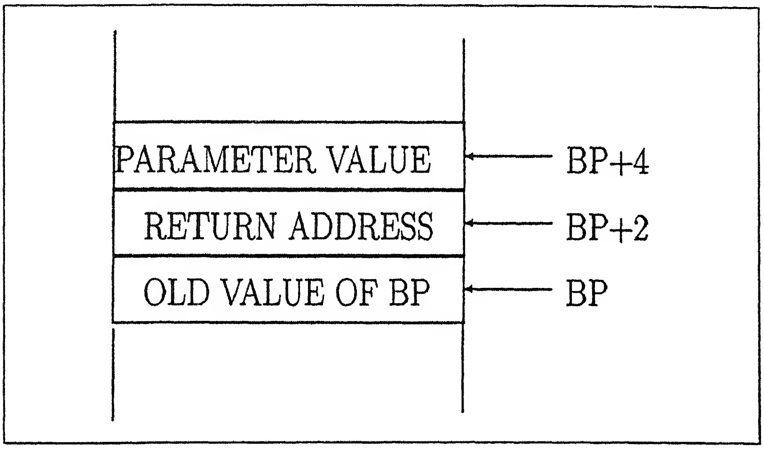


Figure 3 Activation record for factorial computation.

The machine language is nothing but bit-patterns stored in memory. It does not have any comment or labels. (Symbols which denote specific locations in the flow of control in a computer program are called labels. High-level languages as well as assembly language have labels that could be referred to by the programmer. In assembly language, each label is associated with a unique address. Each label is followed by a colon. In the program of *Figure 1*, *enter*, *inside*, *deeper* and *done* are labels. All

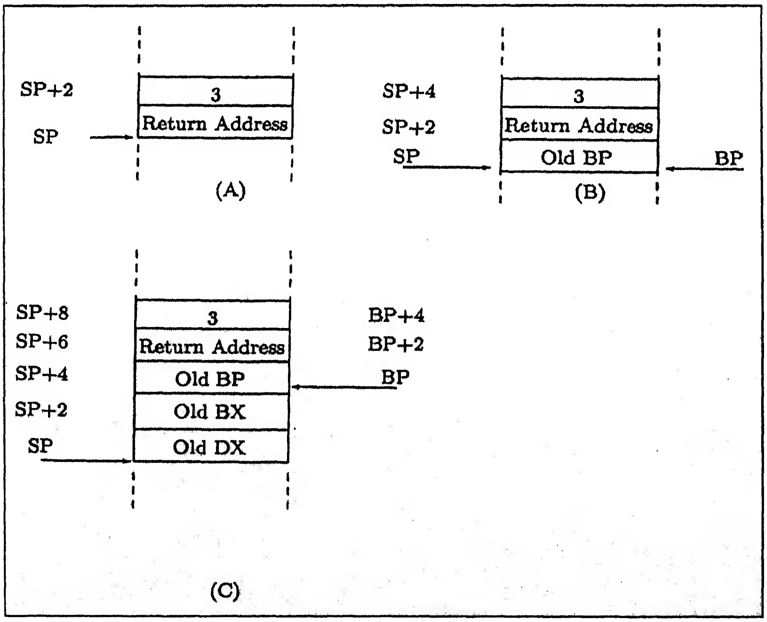


Figure 4 Activation record for n=3.

references to symbolic labels in assembly language are replaced by the actual address which is just a binary number.)

As an example, see how the program of *Figure 1* will look when it is *assembled* or translated into machine language and then *unassembled* again in an attempt to decipher its semantics. *Figure 6* shows a portion of the assembly language program of *Figure 1* along with its machine language object code written as a hexadecimal number. For example, 55 in hexadecimal which is the same bit-pattern as 01010101 in binary is the object code produced from the assembly language statement PUSH BP. And as one can see, it is possible to do a reverse translation from 55 to PUSH BP. In fact, *Figure 6* is generated by unassembling a portion of the machine language program that was generated by assembling the assembly language program of *Figure 1*. However much of the symbolic information like the names of labels (*e.g.* enter: and inside:) and program variables is lost. All comments are lost too. The machine language is designed not to be

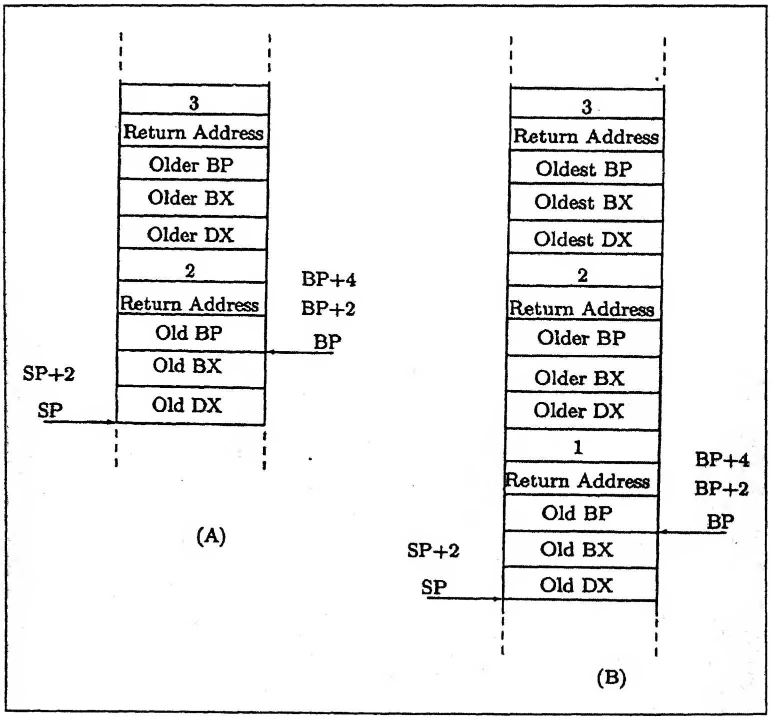


Figure 5 Activation record after recursion proceeds.

understood but to be executed. And to do that, the hardware does not need comments, names or labels. Nor does it need to 'understand' what it is doing. It just does what it is told to do. So the bit-pattern corresponding to the stream of bytes 558BEC5352BA0000 is all that the CPU needs to execute the first five lines of code in *Figure 1.*

However, it is extremely difficult to work with machine language code and figure out its intended purpose, particularly when it is (apparently) not serving this purpose. However, hackers like me must do it when everything else fails.

Let us now see how the program in *Figure 1* works. Remember three rules of parameter passing in C:

- The C caller program pushes the last parameter (or *argument*) first.
- In C, values of arguments are directly pushed into the stack.
- The C caller cleans the stack (this is explained later in this article) once the thread of control returns.

Also remember that in Intel 80X86 architectures, the accumulator is used to return values of C callable functions that return integers. In our case, the factorial function has only one argument and it returns an integer result.

So if the factorial function is called with a parameter value, that value is pushed first into the stack. After that the return address is pushed, as it is a function call and the thread of control must come back to the caller at the C statement, just after the call is

Machine Code	Assembly Language Code
55	push bp
8B EC	mov bp,sp
53	push bx
52	push dx
BA 0000	mov dx,0

Figure 6 Deciphering a machine language program

Assembly Language for all seasons?

Can an assembly language program be written for any problem that can be solved using a computer? Can. this be done if the problem is very complex and is solved using a high-level language? The answer is "Yes". As long as that high-level language is compiled and run as an executable, there exists not only an assembly language program, but also an automatic way of generating that program. It is a different matter altogether that a human programmer will not consider it necessary to write it in assembly language.

Why Should Anyone Ever Write Programs in Assembly Language?

There are C and other high-level language compilers. Programs written in high-level languages are easily portable from one machine to another, easier to understand and explain, easier to modify to serve another purpose and easier to correct (some call this *debugging*) if the program is not working as per its specification. So why should one bother to write a program in assembly language? Many programmers just do. And despite advances made in compiler technology, hardware and high-level language design principles, there are still people who must write some of their programs in assembly language for a good reason. Unlike all other phases of translation, the translation from assembly language to machine language is one-to-one. There is only one machine language program that can be generated from a given assembly language program. This fact makes the execution of assembly language programs predictable and controllable. Thus for all time-critical applications people still use assembly language programming as one can predict the timing of various events within the entire computer system with a resolution of one clock cycle. For controlling hardware and generating different kinds of signals to drive hardware devices, assembly language is indispensable. And all programs become machine language programs and only then run. So all sophisticated applications development environments and programming languages, when the going is smooth, indicate in a highly cultured and refined way that it is doing the job as the human being wanted the job to be done. But when they get into serious trouble, they just crash without being able to report anything at all. Then someone has to use assembly language programming and unassemble the machine language code in order to identify the problem and fix it. (See *Figure 1* of the *Resonance* article in this series on system software). An advanced computing environment is nothing but bare digital hardware surrounded by layers of system software that appears more sophisticated. This is just like a mammal, who, after being trained by different stages of education becomes a computer scientist. And in both cases the background reveals itself from time to time in response to certain basic stimuli. And to recover from such situations, as well as to bring the sophistication back, you need to talk to both of them in a language which their (respective) ancestors knew and carried out the semantic actions without question or comment. Also if you desperately want to know how a given software works, and nobody knows and/or wants to tell you, you can always single step through the machine language code instruction by instruction to find out. For me and many other patient hackers, it often also is the best source of knowledge. There are no two ways to interpret what you read in machine language.

made. The return address is a binary number which indicates the address of the start of the object code produced from the next C statement. The image of the stack is called the *activation record*. It grows in size as more recursive subroutine calls are made, or as subroutines call other subroutines. Each time a subroutine calls itself, a new *instance* of it is created, which has its own context.

So we need a way to distinguish between the contexts of the currently executing subroutine and the contexts of its caller(s) and callee(s). The 80X86 architecture has a special purpose register BP to do just that. The BP points at a location in the stack. The stack grows downwards in the 80X86 architecture. The positions in the stack above what the BP is pointing at belong to the context of the caller. Positions below what the BP is pointing at belong to the context of this subroutine and its callee(s). Thus the value of BP itself at the label *enter*: is the most important key to the context of the subroutine. No matter how deep the recursion is and how huge the context of each of these calls are, given the value of BP, the context can be accessed easily. Also local variables are allocated storage this way. The default storage class which is called automatic in C are local variables. Separate copies (possibly holding different values) of local variables are created whenever the subroutine that defines them is invoked. At runtime, they are allocated storage just by subtracting the total size (in bytes) occupied by them from the SP register, keeping BP as it is. From then on, the local variables can always be referenced at an address a few known byte locations below where the BP is pointing. Registers and other (static) variables, whose values are altered within subroutines are saved inside them and restored at the end by pushing them into the stack and popping them off the stack. Thus the stack pointer SP is a very busy register, whose value changes all the time. Therefore locating any variable by counting up or down from the place where SP is pointing is a nightmare. On the other hand, the value of BP remains unchanged all along the execution of the current instance of the subroutine. That is from *enter*: to *done*: in our case (*Figure 1*) BP keeps pointing at the same place on the stack and all the parameters passed to the factorial subroutine, as well as its own local variables can be easily accessed using BP. So the value of the BP register is the first thing saved in the stack once the thread of control enters a subroutine. That is done by the

PUSH BP instruction in *Figure 1*. After that, the activation record looks like *Figure 3*.

Let us call the assembly language program of *Figure 1* from a C caller with parameter value 3 and profile it. In other words, let us

1. Write a C program as in *Figure 7*.

2. Compile it with a C compiler.

3. Assemble the code of *Figure 1*.

4. Link the two object code modules to produce an executable binary code. Note two points about the linking phase:

(a) There is an underscore in front of the factorial subprogram entry point declaration, whereas there is no such underscore in front of the name of the callee in *Figure 7*. This is no printing mistake. The C compiler puts an underscore in front of all external symbols it generates in the target assembly code. So in order to be recognized by the linker, we deliberately put an underscore in front of the manually written assembly language program.

(b) If you repeat the experiment, put in the appropriate switches to the C compiler and linker so that they use the small memory model ensuring that the virtual address space of the executable program is within 64 KBytes.

5. Run the executable code.

```
main( )
{
 int res;
 res=factorial(3) ;
}
```

Figure 7 A C caller program.

and trace the execution from the point it enters the assembly language procedure at label *_factorial* to the point it goes back to the program in *Figure 7* and assigns a value of 6 to the variable *res*.

The parameter value 3 is pushed first and the return address next. The executable statement that moves the content of the AX register to the variable *res* is what the caller program of *Figure 7* executes on return from the callee. So the address where that statement is loaded is the return address in our case. So when the thread of control enters the callee at label *_factorial,* (see *Figure 1*) the activation record looks like *Figure 4(A)*. On entering the assembly language callee, it first fixes its own context. So the BP of the current instance must copy the value of SP. And before that, the old value of BP must be saved. After that is done, the activation record looks like the diagram in *Figure 4(B)*. BP now stands firm as a rock. The portion of stack above where it is pointing now belongs to the caller and its caller in turn and so on. And the portion below belongs to this instance of *_factorial* and the callee(s) of *_factorial,* if any. From now on until the thread of control executes the ret instruction, SP will change many times. However, BP will stay put, like a bookmark, until the *pop bp* statement just before the *ret*. That *pop bp* will restore the caller's context just before handing control back to the caller. Executing that *pop bp* instruction will destroy this context, but that is exactly what we want.

With BP fixed, we are free to use SP and the stack. Knowing that we need to use the BX register to hold the value of the parameter *n*, we save BX by pushing its old value onto the stack. We also know that the multiply instruction will destroy the DX register. We cannot in any way make the MUL BX instruction spare DX. So we must save DX too. As I pointed out in the last article, this is one more reason why one should not call the registers of the 80X86 'General Purpose Registers'. Just as for DX, every register has a special use which sometimes destroys its contents.

After pushing BX and DX, when the control reaches the label inside:, the stack looks like *Figure 4(C)*. The parameter n is now accessed using BP. It is checked, if n is zero. If the parameter n is zero, then *_factorial* returns a value 1 set in the AX register. Otherwise the recursion goes deeper. The value of n is saved in BX and $n - 1$ is computed in AX. Now *_factorial* imitates its C caller (see *Figure* 7). Just as the caller pushes parameter and then calls *_factorial*, n–1 is pushed into the stack and *_factorial* calls itself. So this whole process is repeated. So for $n = 2$, at *inside*:, the stack looks like *Figure 5(A)* at the label *inside*:. For $n = 1$ look at *Figure 5(B)*. Finally at $n = 0$, the recursion terminates. So the *mul* instruction produces 1*1 for $n = 1$. This returns in AX to the instance of *_factorial* for $n = 2$. This instance produces 1*2 which goes back in AX to the instance for $n = 3$. The mul instruction for this instance produces 2*3. This goes back to the C caller and the variable *res* is set to a value of 6. Note how the assembly language callee *_factorial* cleans the stack after the thread of control returns from its own callee by adding 2 (the size of the parameter pushed in this case) to SP. That must be done, not only to imitate the C caller, but to keep the stack balanced. That is, before a parameter is pushed and a call made, if the value of SP is V_1, then after the thread of control returns, SP must again be made V_1. The act of doing so is called cleaning the stack.

Similarly, if you follow the rules you just learned, you can write assembly language caller programs that call programs written in C. These rules and techniques are fundamental in the implementation of multilayered system software organized in a hierarchical fashion as shown in *Figure 1* of the *Resonance*, April 1996 issue in this series.

Why Recursion Fails in Some Languages

If you write a subroutine in Fortran 77 that calls itself and call it, you will get absurd results. That is because recursion is not supported in Fortran 77 implementation. Parameters are passed using fixed locations in memory and thus one instance overwrites the context of the previous instance(s). Such an implementation cannot distinguish between old, older and oldest like ours did, as shown in *Figure 5.* Language implementations which support recursion, use activation records which grow and shrink dynamically, as was explained in this article. Recursion is so useful that Fortran 90 supports it.

Address for Correspondence
S K Ghoshal
Supercomputer Education
and Research Centre
Indian Institute of Science
Bangalore 560 012, India
email:
ghoshal@serc.iisc.ernet.in
Fax: (080) 334 1683

Suggested Reading

- Allen I Holub. *Compiler Design in C*. Prentice Hall of India Private Limited, 1993.
- V Rajaraman. *Computer Programming in C*. Prentice Hall of India Private Limited, 1995.

Dimensional Analysis

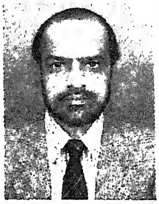
Keeping Track of Length, Mass, Time ....

N N Rao

Dimensional analysis is a useful tool which finds important applications in physics and engineering. It is most effective when there exist a maximal number of dimensionless quantities constructed out of the relevant physical variables. Though a complete theory of dimensional analysis was developed way back in 1914 in a seminal paper by Buckingham, systematic procedures necessary to construct a sufficient number of dimensionless quantities have become available only recently. In this article, we describe with an example the steps involved in the Szirtes algorithm which is fairly simple to understand and quite straightforward to use.

Dimensional analysis is a topic which every student of science encounters in elementary physics courses. The basics of this topic are taught and learnt quite hurriedly (and forgotten fairly quickly thereafter!) It does not generally receive the attention and the respect it deserves even though it has wide applications in all branches of physics and engineering. In the field of physics proper, it is most commonly used as a tool for checking equations for dimensional correctness. (The requirement that all the terms in an equation describing a physical law should have the same dimensions seems to have been first stated by Fourier.) Furthermore, given certain inputs based on experimental observations or data, dimensional analysis can be used to derive empirical laws which would otherwise be quite difficult, if not impossible, to arrive at.

In the field of engineering, dimensional analysis plays a more important role in addition to the above mentioned applications.



Nagesha N Rao is at the Physical Research Laboratory, Ahmedabad. He works in the area of theoretical plasma physics. His recent research interests include theoretical aspects of ionospheric modification experiments, dusty plasma physics and nonlinear dynamics of coupled scalar field equations.
His hobbies include listening to Indian classical music and playing tennis.

Quite often, it is very essential to construct and study a scaled-down model of the system to be investigated or a machine to be constructed.

Quite often, it is very essential to construct and study a scaled-down model of the system to be investigated or a machine to be constructed. For example, prediction of the erosion rate of the river banks would be practically impossible without constructing model structures which properly incorporate the essential qualitative features of a given river system. In aerodynamics, wind turbulence effects on aircraft dynamics is most conveniently studied by using scaled-down models of the full-size systems. In machine design and construction, model studies can significantly reduce the possibility of simple but costly errors in the construction of the final version, generally called the 'prototype'. It is, of course, important that one determines the 'scaling factors' which relate the parameters or the variables of the model to those of the prototype. It is here that the real power of dimensional analysis comes into the picture quite explicitly.

Any important application of dimensional analysis requires the construction of 'dimensionless quantities'. Given a set of physical variables necessary to describe the problem at hand, one constructs by suitable combinations, (by way of multiplications or divisions) quantities where all the dimensions cancel out completely. Dimensionless quantities are constructed and used in almost all the branches of physics and engineering. While the number of such quantities in any given field are generally quite small, there are a few branches of physics where it is almost impossible to discuss any problem without using an appropriate dimensionless quantity. In this connection, it may be interesting to note that fluid mechanics probably tops the list with a total of about 44 dimensionless quantities! As an example of a typical dimensionless quantity, consider the flow of a fluid in a pipe or a channel, which is best described in terms of the fluid densityρ, the pipe radius or the channel width a, the flow velocity$u$ and the fluid viscosity η. Then, the quantity $R \equiv \rho a u/\eta$ (called the Reynolds number) is dimensionless and plays a very important role in determining the transition of the flow from the laminar to the turbulent state. The advantage of using dimensionless quantities is that although their number is generally much smaller than that of the actual physical variables, if chosen properly they would be adequate to characterize certain aspects of system dynamics, such as the flow transition in the example above.

The π-Theorem

A general theory of dimensional analysis and its implications was developed way back in 1914 in a seminal paper by Buckingham. The main result of his work is summarized in a well-known theorem, which is generally referred to as the *π-theorem.* The theorem is applicable to any dimensionally homogeneous equation which relates, say, n physical quantities defined in terms of r reference dimensions (such as M, L and T). A physical equation is said to be dimensionally homogeneous if every term in the equation reduces to the same algebraic quantity when expressed in terms of the reference dimensions. According to the central result of the π-theorem, it is always possible to reduce the equation to a relationship between $(n-r)$ independent dimensionless quantities provided the reference dimensions themselves are considered as independent of one another. The minimal set of such dimensionless quantities for a given system constitutes the fundamental or the complete set. A formal algebraic proof of the Buckingham's π-theorem has been discussed by Isaacson and Isaacson, while Corrsin has given an elegant proof based on geometrical considerations. A number of results of practical importance follow as a consequence of the π-theorem, which have been further discussed in the work done by Bridgman, Ipsen, Duncan and Pankhurst (suggested reading).

The advantage of using dimensionless quantities is that although their number is generally much smaller than that of the actual physical variables, if chosen properly they would be adequate to characterize certain aspects of system dynamics.

While the π-theorem requires the existence of a complete set of dimensionless quantities, there is no unique or universally applicable method to actually construct such quantities explicitly. In his original paper, Buckingham has indicated the basic outline of a procedure that follows as a consequence of the π-theorem. There are, however, practical difficulties in its actual implementation for finding out a minimal set of dimensionless quantities which would not only be appropriate for the problem at hand but also sufficient to describe unambiguously the behavior of the system under consideration. Isaacson and Isaacson have discussed a couple of methods based on the construction of a relevant set of indicial equations. Recently, Thomas Szirtes of SPAR Aerospace, Canada has given a procedure which is better suited for direct application to a number of physical problems, even when the number of physical variables is large. The Szirtes algorithm is fairly simple and quite straightforward, and is formulated in terms of results from the theory of matrices. In the following, I summarize this procedure and illustrate the various steps involved by taking the example of fluid flow in a channel.

The Szirtes Algorithm

According to Szirtes, the following steps are involved in the construction of a minimal set of dimensionless quantities:

Step 1

Given the system, identify the variables, parameters and constants which govern its behavior. It is most essential that all the relevant quantities are included in the list since if any one of them is omitted, the outcome of the dimensional analysis could be erroneous. However, if any irrelevant or superfluous quantity is included, it does not influence the outcome but only makes the analysis more complicated. Thus, when in doubt about the suitability of a quantity, it is best to include it in the list!

As an example, I will consider in the following the flow of a fluid in a pipe or a channel and construct the familiar Reynolds number. Clearly, as mentioned above, the flow would depend on the parameters ρ, a, u, and η. While one can verify *a posteriori* that just these variables are enough to construct the Reynolds number, let us include, say, the surface tension α which may be important in certain flow problems. The number of dimensionless quantities to be constructed would depend on the number of reference

Table 1

	α	ρ	a	u	η
M	1	1	0	0	1
L	0	–3	1	1	–1
T	–2	0	0	–1	–1

dimensions employed. Taking the familiar CGS system with three reference dimensional units (M denoting the mass,L the length and T the time), tabulate the various variables and the units as given in *Table 1.*

The entries in *Table 1* are just the exponents of the corresponding dimensions for each of the variables. For example, viscosity has the dimensions $ML^{-1}T^{-1}$, hence the corresponding values 1, –1, and –1 in the η-column in *Table 1.*

Step 2

Form the square matrix$A$ by taking the right most elements of *Table 1*. Obviously, the order of this matrix would be equal to the number (m) of reference dimensions used. In *Table 1,* we have three reference dimensions, namely,M, L and T and therefore, the matrix A is of the order 3 × 3. Make sure that the matrix A is non-singular, that is, det$A \neq 0$. If det$A = 0$, rearrange the columns in *Table 1* so as to obtain a new square matrix with non-zero determinant. Call the matrix formed by the remaining elements in *Table 1* as matrix$B$ which need not be square.

Thus, for the example of fluid flow, we have

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & -1 \\ 0 & -1 & -1 \end{bmatrix}, \mathrm{B} = \begin{bmatrix} 1 & 1 \\ 0 & -3 \\ -2 & 0 \end{bmatrix}$$

Step 3

Calculate the inverse of the matrix A and find the product matrix C defined by

$$C = -(A^{-1}B)^T ,$$

where T denotes the transpose operation (that is, the interchange of the rows with the corresponding columns) and A^{-1} is the inverse of the matrix A.

For the example under consideration, we can easily show that

$$A^{-1} = \begin{bmatrix} 2 & 1 & 1 \\ -1 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix}, \; C = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 1 & -1 \end{bmatrix}$$

Step 4

Extend *Table 1* as described in the following. Place the matrix C below the matrix A. Place an identity (square) matrix I of appropriate size below the matrix B. Extend the column containing M, L, T by the dimensionless quantities (to be constructed) denoted by, say,π_i, $i=1, 2, 3, ...$ Thus, if the number of physical variables listed is n and the number of reference dimensions used is m with $n > m$, then the dimension of the matrix A is $m\times m$, that of B is $m\times(n-m)$, that of C is $(n-m)\times m$ and that of I is $(n-m)\times(n-m)$. According to the Buckingham π-theorem, the number of dimensionless quantities that can be constructed is given by $N_d=n-m$. Denoting the $m\times 1$ column matrix consisting of the reference dimensions (like$M$, L and T) as the matrix D and the $(n-m)\times 1$ column matrix consisting of the dimensionless quantities (namely, π_1 , π_2 , ..., π_{n-m}) as the matrix π, we have schematically the *Table 2*, where the $1\times n$ row matrix V has as elements the various physical variables (like, α, ρ, a, u, and η) used to describe the system. The dimensions of each of the matrices is shown just below the corresponding entry in *Table 2*. Clearly, the total matrix formed by$A$, B, C, and I is of

Table 2

MATRIX [DIMENSIONS]	V $[1 \times n]$	
D $[m \times 1]$	B $[m \times (n-m)]$	A $[m \times m]$
π $[(n-m) \times 1]$	I $[(n-m) \times (n-m)]$	C $[(n-m) \times m]$

dimension $n \times n$. Each of the dimensionless quantities π_i , $i=1, 2, 3, \ldots, (n-m)$ is then constructed from *Table 2* by forming the product of all the listed physical variables after taking the corresponding entries in the table as their exponents.

For our example, we obtain the output in *Table 3.*

The dimensionless quantities π_1 and π_2 are then easily read off from the *Table 3* as

$$\pi_1 = \alpha^1 \rho^0 a^0 u^{-1} \eta^{-1} = \alpha/u\eta,$$

$$\pi_2 = \alpha^0 \rho^1 a^1 u^1 \eta^{-1} = \rho a u/\eta.$$

Table 3

	α	ρ	a	u	η
M	1	1	0	0	1
L	0	−3	1	1	−1
T	−2	0	0	−1	−1
π_1	1	0	0	−1	−1
π_2	0	1	1	1	−1

Note that the dimensionless quantity π_2 is just the Reynolds number mentioned in the beginning!

It is clear from the above discussion that if the entry α had been omitted in *Table 1*, then the analysis would have yielded only one dimensionless quantity, namely, the Reynolds number (π_2. On the other hand, inclusion of α leads to the existence of an additional variable π_1 which may be useful in certain flow problems where fluid surface tension plays an important and explicit role. Likewise, one could add extra variables to the list which would yield the corresponding dimensionless quantities For example, if the acceleration due to gravity (g) is included, we then obtain a third dimensionless quantity given by $\pi_3 = ag/u^2$

Scaling Laws

The theorem established by Buckingham ensures that the set o dimensionless quantities constructed according to the abov algorithm unambiguously describes the behavior of the physica system under consideration. They can then be used in th construction of models as well as for deriving the scaling laws. A model is said to be dimensionally similar to the prototype if th value of every dimensionless quantity is the same for both. Fo the above example, if the subscripts m and p denote, respectively the corresponding quantities for the model and the prototype for dimensional similarity, we must have

$$\frac{\alpha_m}{u_m \eta_m} = \frac{\alpha_p}{u_p \eta_p}, \quad \frac{\rho_m a_m u_m}{\eta_m} = \frac{\rho_p a_p u_p}{\eta_p},$$

which after rearrangements yield

$$\frac{\alpha_m}{\alpha_p} = \frac{u_m}{u_p} \cdot \frac{\eta_m}{\eta_p}, \quad \frac{\rho_m}{\rho_p} \cdot \frac{a_m}{a_p} = \frac{\eta_m}{\eta_p} \cdot \left(\frac{u_m}{u_p}\right)^{-1}$$

Denoting the various ratios by$S$ (with subscripts corresponding to the various physical variables), we finally obtain the scaling laws,

$$S_\alpha = S_u\, S_\eta\,, \;\; S_\rho S_a = S_\eta S_u^{-1}.$$

These two relations should be satisfied in order that the model and the prototype are dimensionally similar. It should, however, be noted that even though the two systems are dimensionally similar, they could be geometrically quite dissimilar.

Discussion

Let me now describe briefly two uses of dimensionless quantities and the scaling laws mentioned earlier from which one may derive important conclusions about a system and its model. (For illustration, I take the example of fluid flow· discussed above). First, with certain inputs, the qualitative behavior of a given system can be understood easily. For the above example, the input is the experimental fact that beyond a certain value of the Reynolds number (called the critical Reynolds numberR_c which is generally in the range of about 2000) the fluid flow undergoes a transition from the laminar to the turbulent state. Defining the kinematic viscosity $\nu=\eta/\rho$, we have $R=au/\nu$. Therefore, for a given pipe radius a, fluids with larger kinematic viscosity need larger flow speeds in order to undergo the transition. On the other hand, for the same (incompressible) fluid, the onset of turbulent flow occurs at smaller flow speeds for larger pipe radius. Such conclusions which may appear to be obvious can be made quantitative by using the dimensionless quantities. Second, as mentioned above, scaling laws are most important in the construction of models for prototypes. For the above example, if we take the same fluid in the prototype and in the model, then $S_p = 1$, $S_\eta = 1$ and we obtain the simple scaling law $S_aS_u = 1$. Hence, if the scale-size of the model is halved, that is,$S_a = 1/2$, then, for dimensional similarity, the flow speed should be doubled.

Finally, let me conclude by pointing out another important application of dimensionless quantities in a slightly different context. They can be used for simplifying the graphical presentation of experimentally obtained data or theoretically derived equations. Take the example of the Reynolds number. Generally, it is possible to show in one chart in two dimensions, in the simplest way, the functional dependence of three quantities. For example, for a given value of the pipe radius a, we can plot in the $\eta - u$ plane lines of constant density ρ. From the structure of these lines, one can figure out the onset of turbulent flow. However, if such information is needed for a number of different values of a, one requires that many charts. On the other hand, in terms of the Reynolds number R the onset is characterized by just one number, namely, the critical Reynolds number R_c. In fact, if we plot the relation $R = \rho a u/\eta$ in the $R - u$ plane, we obtain a straight line whose slope is just the ratio $\rho a/\eta$. Given the value of R_c, it is then very easy to display graphically the velocity ranges for which the flow is laminar/turbulent and also the transition velocity.

Thus, the topic of 'dimensional analysis' does not seem to be all that trivial!!

An Application to Simulation of Ionospheric 'Heating'

The upper regions of earth's atmosphere at distances of 80 kms and beyond contain matter in a (weakly) ionized state called the *plasma state*. The *ionosphere* not only acts as a protective shell for life on earth but has also played a significant role in making long distance telecommunication a reality. The different layers of the ionosphere act like mirrors for the electromagnetic waves sent from the ground based radio-wave transmitters and reflect them back to the receivers on the ground, thus making 'wireless' communication possible. Needless to say that ionospheric plasma is one of the most intensely studied naturally occurring media surrounding our planet.

Much of our knowledge about the earth's ionosphere has been derived by means of 'passive' experiments carried out over the decades using ground-based, balloon, rocket and, more recently, satellite techniques. On the other hand, with the development of high-power transmitters, it has become possible since the early 1970's to conduct *active* experiments, particularly in the F-region altitudes (150 – 500 kms); that is, as in laboratory experiments, it is possible to induce controlled changes in a small volume of the ionospheric plasma and then study its response to external stimuli, (see Fejer, Leyser and Rao, Kaup in Suggested Reading). Typically, such experiments are carried out using strong electromagnetic (pump) waves in the frequency range ~2 –15 Mhz sent from ground-based transmitters. The pump electromagnetic waves heat the ionospheric plasma in local regions and thereby lead to significant changes in the plasma temperature and number density.

The next natural step is to be able to simulate the ionospheric conditions in laboratory experiments which can be easily controlled. This requires a knowledge of the proper scaling laws between the various variables which characterize the ionospheric system and the simulated model. Hence, it is necessary to identify the relevant dimensionless quantities, and this can be easily done using the Szirtes algorithm as follows:

The ionospheric plasma is characterized by the number density n, the temperature θ, the electron-ion collision frequency ν_{ei} and the ambient terrestrial magnetic field B. The incident pump wave is characterized by the wave frequency ω and the input power P. For a proper sizing of the simulated model, we also need a scale-length variable L. Thus, the matrix V containing the various physical variables which characterize the system is given by,

$$V = [B \quad n \quad \omega \quad P \quad \theta \quad \nu_{ei} \quad L\,] \,.$$

For convenience, let us introduce the temperature θ as an additional reference dimension in addition to M, L and T. We then have the matrices,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 \\ -3 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & -3 & 0 \\ -1 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

- A simple calculation shows that,

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -3 & 0 & -1 & 0 \\ -2 & 1 & 0 & 0 \end{bmatrix}, \quad C =-(A^{-1}B)^T = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & \frac{3}{2} \\ 0 & 0 & 0 & 3 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

Table 2 can now be easily constructed for the present example as,

	B	n	ω	P	θ	ν_{ei}	L
M	$\frac{1}{2}$	0	0	1	0	0	0
L	$-\frac{1}{2}$	-3	0	2	0	0	0
T	-1	0	-1	-3	0	-1	0
θ	0	0	0	0	1	0	0
π_1	1	0	0	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{3}{2}$
π_2	0	1	0	0	0	0	3
π_3	0	0	1	0	0	-1	0

Finally, the three dimensionless quantities π_1, π_2 and π_3 are then easily read off from the above table as,

$$\pi_1 \equiv B^1 P^{-1/2} \nu_{ei}^{1/2} L^{3/2} = \frac{B \nu_{ei}^{1/2} L^{3/2}}{P^{1/2}},$$

$$\pi_2 \equiv n^1 L^3 = nL^3,$$

$$\pi_3 \equiv \omega \nu_{ei}^{-1} = \frac{\omega}{\nu_{ei}},$$

It is easy to recognize the physical content of the quantities π_2 and π_3. The quantity π_2 is simply the total number of plasma particles in a cube of linear dimension L, and is similar to the *plasma parameter*

$(n\ \lambda^3_D)^1$ while π_3 is just the ratio of the wave frequency to the collision frequency. The meaning of π_1 can be made clear by considering

$$\pi_1^2 = \frac{\left(B_0^2\ L^3\right)\nu_{ei}}{P},$$

which is essentially the ratio of the total magnetic field energy per unit collision period to the pump wave power. As earlier, one can now easily derive the scaling laws and thus determine the necessary parameter values for carrying out a scaled down simulation of the ionospheric heating process.

Acknowledgements

I thank K P Subramanian and S K Bhattacharya for useful discussions. I am grateful to the referee for a number of very useful suggestions and constructive comments.

Suggested Reading

- E Buckingham. *Physical Review*. Vol. IV.p 345, 1914.
- S Corrsin. *Americal Journal of Physics*. Vol.19.p 180, 1951.
- W J Duncan. *Physical Similarity and Dimensional Analysis*. Edward Arnold Publishers. London, 1953.
- E C Ipsen. *Units, Dimensions and Dimensionless Numbers*. McGraw-Hill. New York, 1960.
- P W Bridgman. *Dimensional Analysis*. Yale University.Press. New Haven, 1963.
- R C Pankhurst. *Journal of Franklin Institute*. Vol.292.p 451, 1971.
- E de Isaacson and M de Isaacson. *Dimensional Methods in Engineering and Physics*. Edward Arnold Publishers. London, 1975.
- J A Fejer. *Rev. Geophys. Space Phys.* Vol. 17, p.135, 1979.
- T Szirtes. *Machine Design*. p. 113, November 1989.
- T B Leyser and others . *J. Geophys. Res.* Vol.95.p17233, 1990.
- N N Rao and D J Kaup. *J. Geophys. Res.* Vol. 97. p6323, 1992.
- T Szirtes. *SPAR Journal of Engineering and Technology*. Vol. 1. p37, 1992.

Address for Correspondence
N N Rao
Theoretical Physics Division
Physical Research Laboratory
Navrangpura
Ahmedabad 380 009, India

[1] The symbol λ_D is called the *Debye length*. In plasma physics, if one considers a volume much larger than λ_D^3, the Coulomb forces enforce near-equality of the number of positive and negative charges.

Combinatorial Group Theory

Group Theory via Generators and Relations

B Sury

B Sury is with the School of Mathematics, TIFR, Mumbai.

Group theory revolutionized not only mathematics but also other sciences. A combinatorial way of describing groups is by what are called *generators and relations*. In this article, our purpose is to discuss this combinatorial way of describing groups and some of the immediate applications.

Introduction

All of us learn in school, the method of 'completing squares' to solve quadratic equations and perhaps the more precocious ones get to see the formulae for cubic and biquadratic (i.e. fourth degree) equations too. The concept of *groups* surfaced with the fundamental works of the great mathematicians Evariste Galois and Niels Henrik Abel who showed that there is no such *general formula* to solve equations of degree higher than four. It is no exaggeration to say that their ideas revolutionized mathematics and shaped the future direction of algebra. Groups arise in a variety of situations as the group of symmetries of some system. In other words, they arise as the group of transformations of the objects of a system which leave the system as a whole invariant. Fifty years after Galois and Abel, the mathematician Felix Klein introduced his *Erlänger Programm* on the occasion of his admission to the University of Erlangen in 1872, towards a realisation of the fact that any geometry can be characterised by its group of transformations. Sophus Lie, a contemporary of Klein sought to throw light on the solutions of an ordinary differential equation which was invariant under a

The concept of groups surfaced with the fundamental works of the great mathematicians Evariste Galois and Niels Henrik Abel. The advent of group theory revolutionized not only mathematics but also the other sciences.

group of continuous transformations. This theory, known as the theory of Lie groups has applications in numerous branches of mathematics. The advent of group theory revolutionized not only mathematics but also the other sciences. It was not long before it was realized that if groups are studied for their own sake, they would pay heavy dividends. A combinatorial way of describing groups is by what are called 'generators and relations'. This was first developed by W Van Dyck, a student of Klein. In this article, our purpose is to discuss this combinatorial way of describing groups and some of the immediate applications. For unexplained terminology and notation in what is to follow, the reader is encouraged to look up any standard book, such as Herstein's book on elementary algebra (See Suggested Reading).

The fact that all permutations of a set of objects are obtainable from successive interchanges of pairs of objects, can be restated as saying that a symmetric group is generated by its subset of transpositions.

Generators and Relations

A group G is *generated* by a subset S of its elements if every element of G is expressible as a product of elements from S and their inverses i.e. has an expression of the form $g_1^{a_1} \cdots g_k^{a_k}$ with $a_i = \pm 1$ and $g_i \in S$.

For instance, the fact that all permutations of a set of objects are obtainable from successive interchanges of pairs of objects, can be restated as saying that a symmetric group is generated by its subset of transpositions. Obviously, any group has a trivial set of generators viz. itself, but this is hardly of any use; one would like to have a nicer set of generators, preferably a finite set, if one exists. In the latter situation, one calls the group *finitely generated.*

Of course, a finite group is finitely generated. But, there are also several interesting infinite groups that are finitely generated. An obvious example is the additive group $\mathbb{Z}$ of integers; it is generated by the singleton $\{1\}$ (or $\{-1\}$, if you prefer it!)

More generally, for any n, the (so-called) free abelian group of rank n is the additive group $\mathbb{Z}^n := \{(a_1, \cdots, a_n) :$

$a_i \in \mathbb{Z}\}$; this is generated by the usual *basis vectors* $(1,0,\cdots,0),\cdots,(0,0,\cdots,1)$. This is an abelian group[1].

But, there are also infinite nonabelian groups which are finitely generated. For example, look at the subgroup G generated by the matrices $\begin{pmatrix}1 & 2\\0 & 1\end{pmatrix}$ and $\begin{pmatrix}1 & 0\\2 & 1\end{pmatrix}$ inside the group of all 2×2 invertible matrices. G is infinite as the powers $\begin{pmatrix}1 & 2n\\0 & 1\end{pmatrix}$ of $\begin{pmatrix}1 & 2\\0 & 1\end{pmatrix}$ are distinct. G is clearly not abelian since $\begin{pmatrix}1 & 2\\0 & 1\end{pmatrix}$ and $\begin{pmatrix}1 & 0\\2 & 1\end{pmatrix}$ do not commute(Here the group operation is matrix multiplication.).

Note that evidently a group must, at the least, be countable to be finitely generated. Even then, a finite set of generators is not guaranteed.

For instance, the additive group $\mathbb{Q}$ of rational numbers is infinitely generated. For, if $\frac{p_1}{q_1},\ldots,\frac{p_r}{q_r}$ is any finite set in $\mathbb{Q}$, the number $\frac{1}{2q_1\ldots q_r}$ is *not* in the group generated by $\frac{p_1}{q_1},\ldots,\frac{p_r}{q_r}$. (Why?)

Free Groups and the Ping-Pong Lemma

Any (finite) group can be viewed as a subgroup of the group of permutations of a (finite) set. Another way of viewing a group is as a quotient group of a *free group.*

Any (finite) group can be viewed as a subgroup of the group of permutations of a (finite) set. Another way of viewing a group is as a quotient group of a *free* group.

What is a free group? With a given set X of symbols, we first associate a bijective, disjoint set X' of symbols, whose elements will be denoted by x^{-1}. An expression of the form $x_1\cdots x_n$ with $x_i \in X \cup X'$ is called a reduced word, if no x in X appears as a neighbour of x^{-1}. The set of reduced words can be multiplied in a natural way to get a group structure (for $u = x_1\cdots x_n$ and $v = y_1\cdots y_m$, the product $u \cdot v$ is obtained by writing the expression for v after that for u and cancelling off, successively, all pairs of the form xx^{-1} or $x^{-1}x$ occurring as neighbours). We get, then, the *free group $F(X)$ on the set X* where the empty word is the identity element. The cardinality of

[1] The nomenclature abelian is after Abel and is another word for commutative.

X is called the rank of the free group. The rank is an invariant of the group i.e. free groups $F(X), F(Y)$ on sets X, Y are isomorphic if, and only if, the cardinalities of X and Y coincide, and is called the rank of the common group.

Now, for any group G, start with a set X of generators. It is easy to see that G is a quotient group of $F(X)$ i.e. that there is a surjective homomorphism from $F(X)$ onto G. For instance, the group $\mathbb{Z}^n$ of n-tuples of integers mentioned above, is the quotient of the free group F_n of rank n by its commutator subgroup. Recall that the commutator subgroup $[G, G]$ of any group G is defined as the subgroup of G generated by the *commutators* $[x, y] := xyx^{-1}y^{-1}$ of elements in G. Obviously, $\frac{G}{[G,G]}$ is an abelian group; it is called the abelianisation of G.

In the two examples of finitely generated, infinite groups above, the first one of $\mathbb{Z}^n$ is free only for $n = 1$ (as free groups are abelian only in rank one), while the second one is the free group of rank 2, on the two matrices there.

The proof that the matrices $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$ generate a free group is easy and is a consequence of the following trick due to Klein generally known as *the Ping-Pong lemma:*

Suppose there are two nonempty subsets S_1, S_2 and a point p outside them such that two matrices g and h act as transformations on the set $S_1 \cup S_2 \cup \{p\}$ in such a way that $g(S_2 \cup p) \subset S_1$ and $h(S_1 \cup p) \subset S_2$. Then, no nonempty reduced word in g and h acts trivially on p i.e. g and h 'play ping-pong with the point p' between S_1 and S_2 !. Since the group generated by g and h is free precisely when no word in them is the identity word, it would follow that $< g, h >$ (the group generated by g and h)is free.

In our case, we can take $S_1 = \{z \in \mathbb{C} : -1 < Re(z) < 1\}$, $S_2 = \{z \in \mathbb{C} : |z| < 1\}$ and p any point outside the unit circle and with real part between -1 and 1, where the action of any 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is by the fractional linear transformation $z \mapsto \frac{az+b}{cz+d}$. Thus, the two matrices above generate the free group of rank 2.

Presentations of Groups

We talked about generators but we did not say anything about the uniqueness of expressing an element in terms of a generating set. If $g = g_1^{a_1} \cdots g_k^{a_k} = h_1^{b_1} \cdots h_l^{b_l}$ are two different expressions in terms of elements g_i, h_i in a generating set S, there is, obviously, a relation of the form $s_1^{c_1} \cdots s_m^{c_m} = e$ among some elements s_i of S. Clearly, every finite group has relations among any generating set viz. the relation $s^{O(G)} = e$ for any s, where $O(G)$ denotes the number of elements in G.

Let us look at a group G defined by a finite set X of generators and a finite set of relations among them; we write $G =< X; R >$. One has to be careful while talking about *the set of relations.*

If $w = e$ is a relation (where w is a word in the generators), then, so are $w^k = e$ or $xwx^{-1} = e$. But, the latter are consequences of the former. It is in this sense that we say that a set R is a set of defining relations. In the language of free groups, $G =< X; R >$ means that $G \cong F(X)/N$ where N is the subgroup generated by all conjugates[2] of elements of R (i.e. N is called the *normal* subgroup generated in $F(X)$ by the set R). We call $< X; R >$ a *presentation* of G. If both X and R are finite, the group is said to be *finitely presented.* We have to bear in mind that there can be several presentations of the same group.

Let $F = F(X)$ be a free group on a set X. If H is a subgroup of F, it is also free as can be proved naturally by the methods of algebraic topology. It was also proved by combinatorial group-theoretic methods by Nielsen and Schreier. The proofs also show that if X is a finite set of n elements, and if H is of finite index m in F, then the rank of H is $mn - (m - 1)$.

If $G =< X; R >$ is any group with X, R finite sets,

[2] Elements x and y of a group G are said to be conjugate if there exists z in G such that $x = zyz^{-1}$

and H is a subgroup of G of finite index, does H have a finite presentation too, and, if so, how does one find it? Let $\{x_i\}_1^m$ be a set of left-coset representatives for H in G i.e. $G = \cup_1^m x_i H$, a disjoint union. Write $X = \{s_1, \cdots, s_n\}$ and $R = \{w_1, \cdots, w_r\}$ where w_i are words in the s_j. Now $G = F(X)/N$ where N is the normal subgroup of $F(X)$ generated by R. So, $H = E/N$ where $F(X) \supset E \supset N$. Since $[G : H] = [F(X) : E] = m$, E is a free group of rank $mn - (m-1)$ by the Nielsen-Schreier theorem. Thus, H itself is generated by $mn - (m-1)$ elements. It is also obvious that $H = E/M$ where M is the normal subgroup of E generated by the set $\{x_j^{-1} w_i x_j; j \leq m; i \leq r\}$ i.e. H has mr relations.

There is also a beautiful algorithm due to Coxeter, Moser and Todd to write down a presentation for H from one for G. The interested reader might refer to the book *Presentations of groups* by Johnson, published as lecture notes by the London Math. Society (See Suggested Reading).

If G is finitely generated and is also abelian, a fundamental structure theorem of Dedekind says that G is isomorphic to the group

$$\mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_n$$

where the integers d_i divide d_{i+1} and are uniquely determined.

Here we adopt the convention that if $d = 1$, by $\mathbb{Z}/d$ we mean (0), and if $d = 0$, by $\mathbb{Z}/d$ we mean $\mathbb{Z}$. In particular, if G is any finitely generated group, the abelian group $G^{ab} := \frac{G}{[G,G]}$ has the structure asserted above.

The following is a nice way to find the invariants d_i. Let $G =< X; R >$ with $X = \{x_1, \cdots, x_m\}$ and $R = \{w_1, \cdots, w_n\}$. Now, each w_i is a word in the x's. Write M for the $m \times n$ integer matrix whose (i,j)-th entry m_{ij} is the sum of the powers of x_i occurring in the expression of w_j. Let $h_i(M)$ denote the G.C.D of all the $i \times i$ minors of M, for $i \leq k := min(m,n)$. Let $d_1 = h_1(M)$ and $d_i(M) = \frac{h_i(M)}{h_{i-1}(M)}$ $\forall i > 1$. Then, the invariants of $\frac{G}{[G,G]}$ are $d_1, \cdots, d_k, 0, 0, \ldots, 0$ where $k = min(m,n)$ and 0 is repeated $m-k$ times. In other words, $\frac{G}{[G,G]} \cong \mathbb{Z}^{m-k} \times \mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_k$. In particular, we notice that if $m > n$, then $m > k$ and so, $\frac{G}{[G,G]}$ is infinite.

This shows also that if $G =< X; R >$ is a *finite* group, then $m \leq n$ i.e. *the number of generators in any presentation of a finite group is at the most the number of relations* !

What can one say about a finitely generated group where each element has finite order? Is such a group necessarily finite? This is the famous Burnside problem and the answer is negative even when the orders of all the elements are bounded by a fixed number.

The Burnside Problem

What can one say about a finitely generated group where each element has finite order? Is such a group necessarily finite? This is the famous Burnside problem [3] and the answer is negative even when the orders of all the elements are bounded by a fixed number. However, it is very difficult to give such an example of a finitely generated, infinite group all of whose elements are of orders less than some fixed r.

But, there are positive results too. For instance, a group all of whose (nontrivial) elements are of order 2 is clearly abelian (because $x^2 = 1$ for all x means $x = x^{-1}$ i.e. $ab = (ab)^{-1} = b^{-1}a^{-1} = ba$). So, if such a group were finitely generated also, then all the elements of the group are found among the finite set $x_1^{\pm 1} \cdots x_n^{\pm 1}$ where $x_1, \cdots, x_n$ is a set of generators for G.

Even those finitely generated groups all of whose elements have order ≤ 3 are necessarily finite (although nonabelian in general).

Consider a finitely generated group G consisting of matrices with entries in the complex field. If all matrices in G have finite orders, G is necessarily finite. The proof uses some sophisticated methods and we don't comment on it here. On the other hand, if the entries of the matrices in G are integers, instead of complex numbers, the proof is quite easy as we show now.

Note first that since $\det(g)$ and $\det(g^{-1}) = (\det\,(g))^{-1}$ are

[3] Work on problems related to this earned E I Zelmanov the Fields Medal in 1994.

both integers, we have $\det(g) = \pm 1$ for all $g \in G$. Call $GL(\mathrm{n}, \mathbb{Z}) := \{g : g_{ij} \in \mathbb{Z}, det(g) = \pm 1\}$.[4] So, our group G is a subgroup of $GL(\mathrm{n}, \mathbb{Z})$. Observe that if $g \in GL(n, \mathbb{Z})$ has order r, its eigen values are r-th roots of unity and, so, the minimal polynomial $P(X)$ of g divides the polynomial $X^r - 1$. As a result, it has distinct roots.; so g can be conjugated (by some complex matrix) to a diagonal matrix $\mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ where λ_i are r-th roots of 1. (Why ?) Thus, the trace of g satisfies

$$| Tr(g) |=| \sum \lambda_i |\leq \sum | \lambda_i |= n$$

As $Tr(g)$ is an integer, the condition $| Tr(g) |\leq n$ implies that the possible values of $Tr(g)$ are among $\{n, n-1, \cdots, 0, -1, \cdots, -n\}$. To summarise, *any matrix of finite order* in $GL(n, \mathbb{Z})$ has trace in the finite set $\{0, \pm 1, \cdots, \pm n\}$. Let p be a prime not dividing $(2n)!$. Consider the finite group $GL(n, \mathbb{Z}/p)$ of $n \times n$ invertible matrices with entries in $\mathbb{Z}/p$. Look at the natural homomorphism obtained by reducing each entry mod p

$$\phi : GL(n, \mathbb{Z}) \rightarrow GL(n, \mathbb{Z}/p)$$

We claim that $G \bigcap Ker(\phi) = \{Id\}$ i.e. that $G \cong \phi(G)$. If $g \in Ker(\phi)$ has finite order, then $\overline{g_{ij}} = \overline{\delta_{ij}}$, where the 'bar' denotes mod p. So, $Tr(g) \equiv n \bmod p$. But $Tr(g) - n$ takes values in $\{0, -1, \cdots, -2n\}$ since $Tr(g)$ takes values in the set $\{0, \pm 1, \cdots, \pm n\}$. By the choice of p, this forces $Tr(g) = n$ i.e. $g = Id$. So, $\phi(G) \cong G$ and, therefore, $| G |\leq| GL(n, \mathbb{Z}/p) |$.

Probability of Generating the Integers

We end with a *heuristic* discussion which can be made rigorous.

What is the 'probability' P that two randomly chosen integers generate $\mathbb{Z}$? Well, they generate *some* subgroup of $\mathbb{Z}$, at any rate. Overlooking the case that this subgroup is $\{0\}$ (surely an event of probability 0), this subgroup is

[4] Thus $GL(n, \mathbb{Z})$ is the collection of all $n \times n$ matrices with integer entries and determinant either +1 or −1.

of $\mathbb{Z}$, at any rate. Overlooking the case that this subgroup is $\{0\}$ (surely an event of probability 0), this subgroup is $n\mathbb{Z}$ for some $n > 0$. The probability that both the integers belong to $n\mathbb{Z}$ is $\frac{1}{n^2}$. Since $n\mathbb{Z} \cong \mathbb{Z}$, P is also the probability that two elements of $n\mathbb{Z}$ generate it;' and so $\frac{P}{n^2}$ is the probability that two random integers generate $n\mathbb{Z}$. Therefore, $\sum_{n=1}^{\infty} \frac{P}{n^2} = 1$ which gives $P = \frac{1}{\sum_{n=1}^{\infty} \frac{1}{n^2}}$.

On the other hand, two integers generate $\mathbb{Z}$ exactly when they are coprime. Since every pth integer is divisible by p, the 'probability' of a 'randomly' chosen integer being divisible by p can be taken to be $\frac{1}{p}$. Thus the probability that two independently chosen integers are both divisible by p is $\frac{1}{p^2}$; hence the probability that not both are multiples of p is $1 - \cdot \frac{1}{p^2}$. Therefore, the probability that they are coprime is the probability that not both of them are multiples of any prime i.e. $P = \prod_{p\ prime}(1 - \frac{1}{p^2})$. We, thereby, get the 'Euler product formula'

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \prod_{p\ prime} (1 - \frac{1}{p^2})^{-1}$$

However, we know that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$, so that we finally have $P = \frac{6}{\pi^2}$.

A similar discussion can be made for any positive integer k in place of two integers. For $k = 1$, this probability is obviously 0 (as only ± 1 generate $\mathbb{Z}$); and is also $\prod_{p\ prime}(1 - \frac{1}{p})$, on the other hand. This shows that $\prod_{p\ prime}(1 - \frac{1}{p})^{-1}$ diverges i.e. the number of primes is infinite.

The discussion above was not rigorous as probability was not defined precisely. All of this can be done precisely, in terms of the notion of the *Haar measure* on a *profinite group* and (hopefully) this will be done in a follow-up article!

Suggested Reading

- I N Herstein. *Topics in Algebra.* Second edition. Vikas Publishing House, 1976.

For the uninitiated reader, who wants to look up standard notation and terminology used in elementary group theory, this book is a good source.

- D L Johnson. *Presentation of groups.* London Math. Soc. Lecture Note Series. No.22. Cambridge University Press.

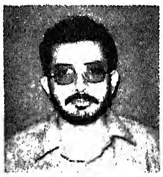
Address for Correspondence
B Sury
School of Mathematics
Tata Institute of Fundamental
Research, Homi Bhabha Road
Mumbai 400 005, India
email: sury@tifrvax.tifr.res.in

Evolution, Fruit Flies and Gerontology

Evolutionary Biology Helps Unravel the Mysteries of Ageing

Amitabh Joshi

In the past decade or so, genetic theories of the evolution of ageing and studies on populations of fruit flies (*Drosophila* spp.) in the laboratory have provided a new perspective on the phenomenon of ageing. These recent advances, very different in approach and methodology from traditional gerontological studies, have provided a wealth of knowledge about the mechanisms of ageing, as well as some answers to deeper, more philosophical questions, such as "why do organisms age at all?".



Amitabh Joshi studies evolutionary genetics at the Jawaharlal Nehru Centre for Advanced Scientific Research. He obtained his Ph.D. from Washington State University in 1993, working on the evolutionary ecology and genetics of species interactions under John N Thompson. Subsequently, he was at the Univ. of California, Irvine, studying ageing and other aspects of life-history evolution in fruit flies.

Different Approaches to the Issue of Ageing

Since earliest times, people have been aware of, and indeed obsessed with, the phenomenon of ageing. Mythology and history abound with examples of mankind's quest for a formula that would either ensure immortality, or at least postpone the general decline in health and well-being associated with ageing. By the nineteenth century, this interest had crystallised into a well defined approach followed by scientists, many of whom were medical doctors, who were interested in understanding the processes of human ageing with the ultimate goal of postponing ageing in humans. This approach, rooted in mammalian physiology, gave rise to modern gerontology, which is still characterised by a focus on physiological changes accompanying human (or at least mammalian) ageing.

On the other hand, nineteenth century thinkers like Alfred Russell Wallace (co-propounder, with Darwin, of the theory of evolution by natural selection) and August Weismann had already begun to speculate on the mechanisms causing the evolution of ageing as opposed to the actual mechanism of ageing in any particular species. However, lacking the foundation of modern evolutionary genetics, their theories tended to lapse into vague idealistic arguments. They thought of ageing as being for the long-term good of the species, rather than the immediate benefit of the individual. Consequently, it was not until the middle of the twentieth century, after the development of population genetic theory, that evolutionary theories of ageing were clearly enunciated by scientists like J B S Haldane and Sir Peter Medawar.

The principal difference between the gerontological and evolutionary approaches to the study of ageing lies in the nature of the questions asked, even though both approaches attempt to unravel the mechanisms of ageing.

The principal difference between the gerontological and evolutionary approaches to the study of ageing lies in the nature of the questions asked, even though both approaches attempt to unravel the mechanisms of ageing. Gerontologists tend to focus on what are called *proximal causes* (*e.g.* looking at the effect of hormone levels on ageing), whereas the evolutionary biologist focuses on so-called *ultimate causes* (*e.g.* why ageing has evolved in certain species and not in others). As we shall see, the latter approach has ultimately led to a better understanding of proximal causes of ageing as well.

What is Ageing?

At this juncture, it would be useful to point out that not all organisms age (can you think of any in light of the following definition of ageing?), and that the word ageing (or senescence) has a fairly precise meaning for biologists. A decrease in the likelihood of surviving with increasing age is not necessarily ageing. Even if the probability of death at any point in time, say from accidents, predation, or just wear and tear, is constant (or even decreasing), it will result in a decreasing probability of survival with age (see *Box 1*). A definition of ageing commonly used by evolutionary biologists and demographers is that *ageing is an acceleration in the rate at which the probability of survival declines with age, relative to the rate associated with juveniles measured under conditions free of externally imposed sources of mortality.*

Ageing is an acceleration in the rate at which the probability of survival declines with age, relative to the rate associated with juveniles measured under conditions free of externally imposed sources of mortality.

Mortality Rates and Survival Probability

To see why a constant mortality rate (*i.e.* a constant probability of death during a given , or interval, of time) gives rise to a decrease in the likelihood of survival with increasing age, consider the following example. In your kitchen you have ten beautiful porcelain cups. You also have a very clumsy servant, whose job it is to wash these cups each day. Let us say that the probability that the servant breaks a cup while washing it is 1/10. Intuitively, one would imagine that after a few days you would have very few cups left, and that ultimately all the cups would get broken. In other words, the fraction of cups still intact at any given day will tend to decrease over time. This empirical observation about the decreasing proportion, over time, of intact cups can also be expressed as a decrease in the probability that a cup is still intact some arbitrary number of days after the point at which all ten were intact. The critical point in this example is that even though the probability of a cup getting broken each day is constant from day to day, the probability of a cup being intact on a given day is one minus the cumulative probability that it got broken on *any of the preceding days.* The cumulative probability of getting broken tends to increase with the passage of time, causing a concomitant decline in the probability of still being intact. The reader can, no doubt, generalise from cups to organisms and from clumsy servants to the vagaries of fate.

Evolutionary Theories of Ageing

There are two major genetic theories that suggest mechanisms by which the phenomenon of ageing might have evolved: antagonistic pleiotropy and mutation accumulation. Both these theories build upon a basic tenet of evolutionary biology, namely that the strength of natural selection acting on a gene tends to decrease with the age at which the gene is expressed in the organisms carrying it (see*Box 2*). In genetics, pleiotropy refers to a situation where one gene affects two or more traits. Antagonistic pleiotropy occurs when the effects of the same gene on two different traits are opposite in nature. For example, if a particular gene increased your height but simultaneously decreased your weight, it would be considered to show antagonistic pleiotropic effects on height and weight. One especially well documented case of antagonistic pleiotropy involves egg production and life-span. These two traits are inversely related in many organisms

Pleiotropy refers to a situation where one gene affects two or more traits. Antagonistic pleiotropy occurs when the effects of the same gene on two different traits are opposite in nature.

The Force of Natural Selection Declines with Age

In evolutionary biology, natural selection is often likened to a force that tends to change the genetic composition of a population over generations, thereby leading to adaptive evolution. A simple, but important, tenet of evolutionary theory is that the strength of natural selection acting on a gene tends to decrease with the age at which the gene is expressed in the organisms carrying it. As an example, consider two genes in humans, one of which kills its carriers around the age of 15 while the other kills its carriers around the age of 40. Clearly, the first gene will rather rapidly disappear from a population because individuals carrying it die before they are likely to have had children; the gene is not likely to be passed on to subsequent generations and is, thus, being eliminated by natural selection. The gene that kills its carriers around the age of 40, however, will be transmitted to the next generation almost as efficiently as a gene that has no harmful effect at all, because by 40 most people have had as many children as they are likely to have for life. Thus, even though both genes are equally harmful in a physiological sense (both are lethal), one escapes natural selection because its harmful effect is expressed only late in life at a time when it no longer really affects the number of offspring its carriers can bear. Both major genetic theories of the evolution of ageing (antagonistic pleiotropy and mutation accumulation) are based upon the fact that the strength of natural selection declines with age.

including *Drosophila*. Females that produce more offspring early in life tend to die earlier than those that produce less offspring. The genes resulting in the greater egg production of such females can, therefore, be considered to simultaneously reduce their life-span. These genes may be thought of as increasing the Darwinian fitness (*i.e.* the ability to contribute offspring to subsequent generations) of their carriers early in life because they enable the carriers to lay more eggs. At the same time, by reducing life-span, these genes are also reducing the fitness of their carriers, but only relatively late in life. Evolutionary theory predicts that genes that increase fitness early in life will be favoured by natural selection, even if they have harmful effects later on. This is a simple consequence of the fact that the force of natural selection declines with the age at which a gene is expressed. Selection against a gene that is harmful later in life will be relatively weak and will, therefore, tend to be overshadowed by selection in favour of the gene based on its beneficial effects early on. Consequently, the proportion of genes with beneficial effects

The force of natural selection declines with the age at which a gene is expressed.

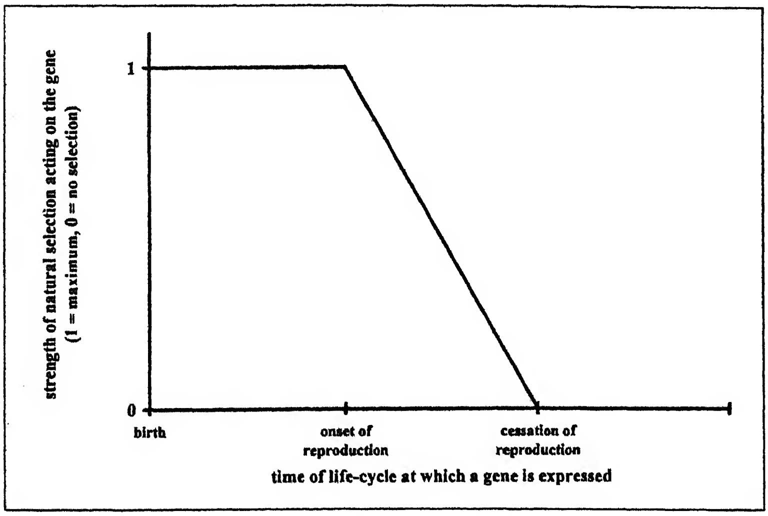


Figure 1 The force of natural selection declines with age. The strength of natural selection is at a maximum when acting on genes expressed before the onset of reproduction. It declines with respect to genes expressed at later and later ages during the reproductive phase of the organism's life-cycle. Genes expressed after the cessation of reproduction are not acted upon by natural selection at all. The steepness of the decline in the strength of natural selection during the reproductive phase increases when reproductive events tend to be more frequent earlier in the reproductive phase of the life-cycle.

early in life and harmful effects later on will tend to increase over generations in populations. In this view, then, ageing emerges as the manifestation late in life of the harmful pleiotropic effects of these genes, almost a by-product of natural selection for enhanced reproductive ability early in life.

The mutation accumulation theory of the evolution of ageing is very similar conceptually to that of antagonistic pleiotropy. In this theory, it is assumed that harmful mutant genes continually arise in populations. Some of these mutant genes exert their harmful effects early in life. Such genes are eliminated from populations by natural selection. Other mutant genes have no effects early on, but have harmful effects relatively late in life. Due to the declining force of natural selection with age, these genes escape elimination by selection and persist in populations. Some such genes may even become more common in a population over many generations entirely due to chance (a phenomenon referred to as random genetic drift). Thus, over a long span of time, many such genes with no effects early in life, but harmful effects later on, may accumulate in populations and contribute to the visible deterioration associated with ageing, because most individuals in such populations will be carrying a number of these genes that are harmful late in life. Note that the mechanisms invoked by the theories of mutation accumulation and antagonistic pleiotropy are not mutually exclusive: both phenomena could well be occurring together in populations.

The antagonistic pleiotropy theory assumes that some genes are beneficial early on but harmful at later ages.

Both the theories described above are based upon the decline in the force of natural selection with age. The principal differences between the two theories are in the types of genes causing ageing that are assumed to occur in populations, and in the evolutionary mechanism by which these genes become common in populations over the course of evolution. The antagonistic pleiotropy theory assumes that some genes are beneficial early on but harmful at later ages. Such genes become more common in populations over time through natural selection. In the mutation accumulation scenario, genes harmful later in life which, however, have no effects on fitness early on, become more common over time through the random process of genetic drift.

Testing Evolutionary Theories of Ageing with *Drosophila*

There is now empirical evidence demonstrating the occurrence of both these mechanisms of ageing in laboratory populations of the fruit fly *Drosophila melanogaster.*

Both of the evolutionary theories of ageing give rise to predictions that can be tested in the laboratory in a fairly straightforward manner. Indeed, there is now empirical evidence demonstrating the occurrence of both these mechanisms of ageing in laboratory populations of the fruit fly *Drosophila melanogaster.* If the antagonistic pleiotropy theory holds true, then genes increasing some aspect of fitness (*e.g.* egg production) early in life should have detrimental effects later on in life, and vice versa. By subjecting fruit flies to natural selection in the laboratory, Michael Rose (a Canadian evolutionary biologist, now at University of California, Irvine) and his colleagues were able to confirm this prediction (see *Box 3* for details on how this was done). Populations that evolved increased longevity and egg production late in life, as a response to selection, showed a correlated decline in their egg production early in life. On the other hand, populations that initially had high longevity and relatively low egg production early in life underwent a correlated decrease in

Natural Selection for Increased Life-Span in the Laboratory

In an evolutionary sense, selection is said to occur when different genotypes tend to produce different numbers of offspring, thereby resulting in some genotypes (the more fit ones) being better represented in the next generation than others (the less fit ones). In natural selection, whether in the wild or in a laboratory, the relative fitness of genotypes is determined entirely by how well adapted they happen to be to their environment. In artificial selection, as practised by plant and animal breeders, the genotypes of individuals are inferred from their phenotypes and the experimenter then determines which of those genotypes will be permitted to contribute offspring to the next generation. To select for increased life-span in the laboratory, Michael Rose and his colleagues started with five populations of *D. melanogaster* that had a mean life-span of about 35 days. In these populations, adult flies emerged from the pupae at about 9 days of age. The eggs laid by the flies when they were 14 days old were then used to start the next generation, and the adults were discarded. From these populations, Rose and his colleagues derived five new populations in which they pushed forward each generation the day on which eggs were collected for initiating the next generation. For example, in the first generation of selection, they may have collected eggs on day 19 rather than 14, and in the second generation, at day 24 rather than 19. By changing the day of egg collection, they ensured that those genotypes that happened to be better able to produce eggs at a slightly older age would be better represented in each subsequent generation: natural selection for increased egg-production later in life, which implies indirect selection for life-span, because to lay more eggs late in life an individual first must survive that long (Incidentally, can you think of how artificial selection for increased life-span would be done?).

longevity and late-life egg production when they were successfully subjected to selection for increased egg-laying early in life. This symmetric pattern of evolutionary change in opposite directions for early and late life fitness traits, respectively, provides experimental support for the antagonistic pleiotropy theory.

Under the mutation accumulation scenario, different populations are expected to harbour at least partly different sets of genes that have harmful effects later in life, as a consequence of the random nature of the accumulation of such genes in populations through random genetic drift. Given that most harmful mutations tend to be recessive (why that is so is a rather interesting issue but, unfortunately, beyond the scope of this article), one would expect that if two populations that exhibited ageing because of mutation

Given that most harmful mutations tend to be recessive, one would expect that if two populations which exhibited ageing because of mutation accumulation were crossed, then the hybrids would perform better than the parents at advanced ages.

accumulation were crossed, then the hybrids would perform better than the parents at advanced ages (see *Box 4*). This is because the hybrids would be heterozygous for at least some of the genes with harmful effects late in life that existed in the parental populations. Such an effect was observed by Laurence Mueller (now at University of California, Irvine) when he carried out all possible pair-wise crosses among three populations of *D. melanogaster* that showed a senescent decline in egg production from an age of about 3 weeks onward. The hybrids consistently produced more eggs than their parents at ages of 3 or 4 weeks, suggesting that the senescent decline in the parental populations was, indeed, a consequence of mutation accumulation. The fact that the hybrids did not lay more eggs than their parents at earlier ages rules out the possibility that the result was merely an expression of *hybrid vigour*, a situation where hybrids outperform their parents at all ages.

Drosophila with Postponed Ageing as a Model System

The five long-lived populations of flies have provided what is perhaps the best system in the world for studying the behavioural, physiological, biochemical and genetic correlates of the ageing process.

We have seen earlier that in order to test evolutionary theories of ageing, Rose and his colleagues created extremely long-lived flies in the laboratory by subjecting them to natural selection for increased life-span and egg-laying at advanced ages. These five long-lived populations of flies have since provided what is perhaps the best system in the world for studying the behavioural, physiological, biochemical and genetic correlates of the ageing process. The reason for this is that these flies live up to three or four times longer than normal *D. melanogaster* would under laboratory conditions. These long-lived flies have an average life-span exceeding 120 days whereas flies from control populations typically live about 35 days. Comparison of the long-lived flies with control flies, therefore, allows the identification of those aspects of their biology that differ from the control flies, and are, consequently, likely to play a major role in ageing. The Ageing Group at University of California, Irvine, headed by Rose, has been studying these populations intensively for the last

Testing the Mutation Accumulation Theory

To see why crossing two populations and comparing the hybrids to their parents allows us to test the mutation accumulation theory of the evolution of ageing, consider the following scenario. We have two populations (lets call them X and Y) that show some pattern of ageing, say a decline in egg production at the age of 3 weeks. We assume that this is because certain genes that adversely affect egg production at week 3, but not before, have become very common in these populations through random genetic drift (this is just another way of saying that we assume that the pattern of ageing in these populations has evolved through mutation accumulation). Now, because genetic drift is a random process, we assume that at least some of these genes are unique to one or the other population. There may be other genes that have become common in both populations by chance, but that really does not make any difference to the argument. Given that most harmful mutations tend to be recessive, we can then write out a partial genotype for an average individual from each of these two populations as follows.

Average individual from population X: *AAbbCCdd*

Average individual from population Y: *AABBccdd*

Here, at the A locus, both populations lack the harmful allele *a*, whereas at the D locus, the harmful recessive *d* allele has become very common in both populations. The crucial loci for testing mutation accumulation, however, are the B and C loci. In population X, the harmful recessive allele *b* has become very common (that is why the average member of this population has the genotype *bb* at this locus). In population Y, on the other hand, the harmful recessive allele *c* has become very common. The four loci considered in this example represent four categories of loci that may be thought to occur in populations. The actual number of loci of each type will, of course, vary from one population to the next. All that is required for testing the theory of mutation accumulation is that there be at least some loci in the two populations being studied that show the pattern depicted here for the B and C loci. Now consider the hybrid arising from a cross of individuals from the X and Y populations.

$$AAbbCCdd \times AABBccdd \rightarrow AABbCcdd$$

The hybrids are heterozygous at the B and C loci. Because the harmful alleles *b* and *c* are recessive, these hybrids will escape their harmful effects. Consequently, at the third week the hybrids should be able to lay more eggs than flies from the parental populations X and Y. If such a pattern is indeed seen, we conclude that mutation accumulation was responsible for the pattern of ageing seen in populations X and Y. Had antagonistic pleiotropy been the sole cause of the observed pattern of ageing, we would not expect the hybrids to lay more eggs than their parents at week 3. This is because under the antagonistic pleiotropy scenario the genes causing ageing become common in populations by the directional force of natural selection rather than by a random mechanism such as drift. Therefore, we would expect the same genes to become common in both populations. The hybrids would then remain homozygous for those genes because both parents would be homozygous at those loci.

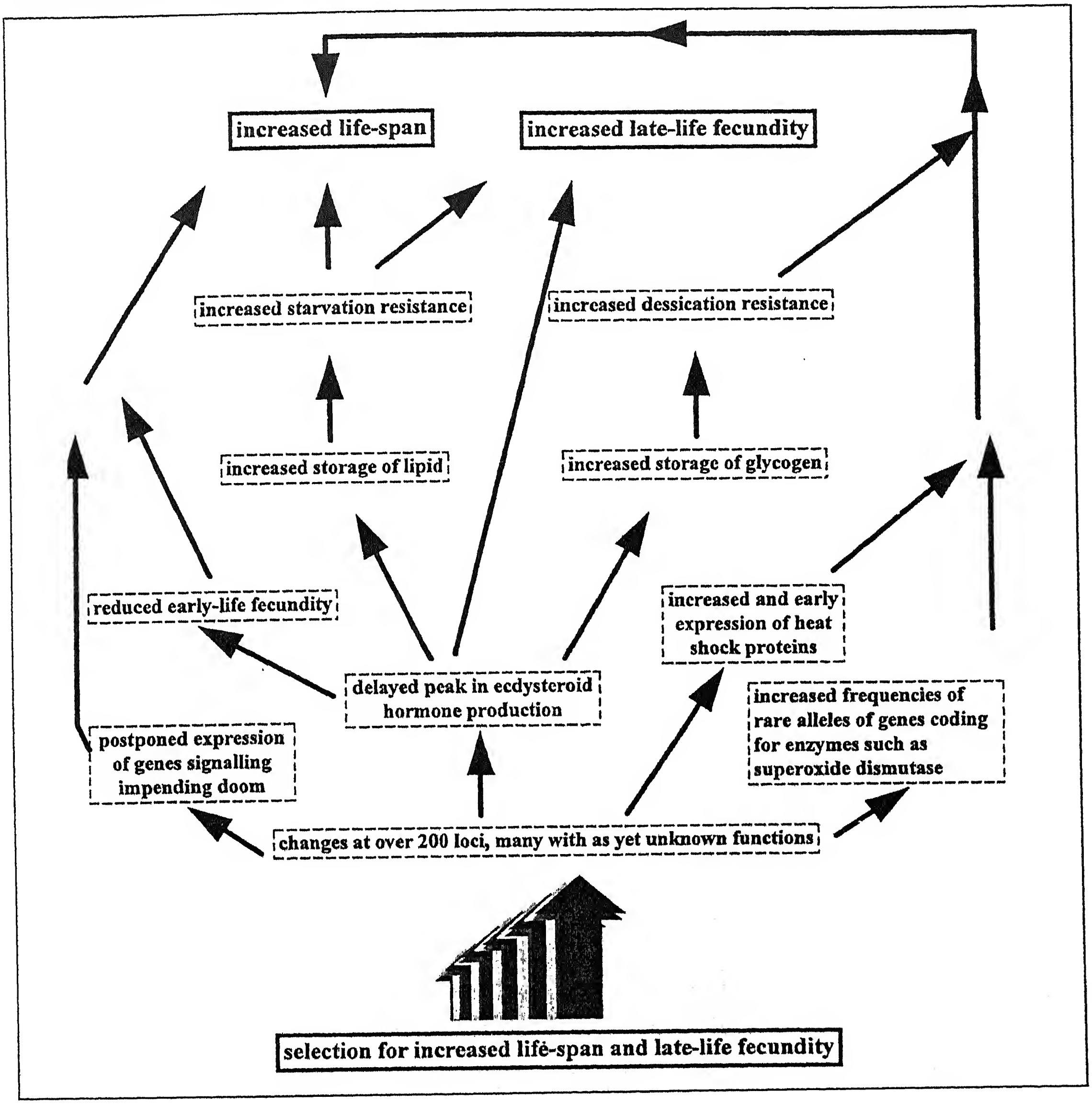
several years. The group is a diverse collection of scientists with backgrounds in evolutionary biology, insect physiology, endocrinology and molecular genetics.

One of the principal early findings of this group was that their long-lived flies tended to have higher resistance to various stresses like starvation and desiccation.

One of the principal early findings of this group was that their long-lived flies tended to have higher resistance to various stresses like starvation and desiccation. This was also confirmed by selecting flies for increased stress resistance, and observing that such flies also evolved a greater life-span. Moreover, the greater stress resistance of the long-lived flies was shown to be largely due to increased accumulation of food reserves, especially lipids, but also glycogen which acts as a reservoir for water in the body. Differences between long-lived and control flies in patterns of lipid storage pointed to the involvement of hormones such as ecdysone in the ageing process. These hormones are typically activated by mating and, among other things, cause a mobilisation of stored lipid reserves for egg production. As one might expect, the level of ecdysone-like hormones in young flies of the long-lived populations turned out to be considerably lower than in control flies of the same age. Electrophoretic studies have also revealed differences between long-lived and control flies in the relative frequencies of genes coding for alternative forms of several enzymes like superoxide dismutase (the rare S allele, coding for the more active form of the enzyme, has risen to high frequency in the O populations) and phosphoglucomutase. Superoxide dismutase has long been thought to be·linked to ageing because of its role, along with catalase, in preventing oxidative damage to cells by free radicals, and these observations lend empirical credence to this view. More detailed biochemical and genetic studies on these populations are currently underway to further elucidate the role of hormones and specific genes in the process of ageing. The strengths of this system are that direct comparisons can now be made, among populations of the same species, of individuals with normal and delayed patterns of ageing. This is a tremendous advance over the traditional gerontological approach of looking at small numbers of highly inbred individuals with accelerated ageing. A major problem

Differences between long-lived and control flies in patterns of lipid storage pointed to the involvement of hormones such as ecdysone in the ageing process.

Figure 2 Multiple factors causing increased life-span in fruit flies. A schematic representation of the cascade of events underlying the successful response to selection for increased life-span in D. melanogaster. Many of the pathways shown are tentative and based on suggestive but, nevertheless, preliminary results. Many others are, undoubtedly, yet to be discovered.


increased life-span
increased late-life fecundity
increased starvation resistance
increased dessication resistance
increased storage of lipid
increased storage of glycogen
reduced early-life fecundity
increased and early expression of heat shock proteins
delayed peak in ecdysteroid hormone production
increased frequencies of rare alleles of genes coding for enzymes such as superoxide dismutase
postponed expression of genes signalling impending doom
changes at over 200 loci, many with as yet unknown functions
selection for increased life-span and late-life fecundity

with that approach was the inability to distinguish pathological effects causing early death from a true change in the rate of ageing. Moreover, conclusions drawn from highly inbred organisms are not easily extended to species showing crossbreeding, which is a much more common situation in nature.

Conclusion

Searching for a single underlying physiological cause of ageing across many species may well prove unfruitful.

The experimental validation of the evolutionary theories of ageing has some interesting implications for gerontology. Searching for a single underlying physiological cause of ageing across many species may well prove unfruitful. Given that ageing has evolved by the genetic mechanisms outlined above, there is no particular reason to expect physiological uniformity in ageing across species because it is unlikely that the same kinds of harmful genes would have become common in various species over the course of their evolution. Similarly, even within a species, there is likely to be a multitude of physiological processes contributing to the senescent decline associated with ageing. Indeed, in the words of Michael Rose, "*The only universal mechanism involved in senescence is that of its evolution*". Moreover, the evolutionary approach to ageing suggests that there is nothing about ageing that is in some sense intrinsic and fundamental to life itself. In the evolutionary theories, ageing, in fact, is seen to be a by-product of the way in which natural selection acts on organisms with certain types of life-histories. To answer a question raised earlier in this article, single-celled organisms do not age, neither do asexually propagating organisms such as corals and many plants. Moreover, organisms that die almost immediately after their first round of reproduction, such as annual plants and many insects, are also not considered to undergo ageing. Ageing, thus, is seen to be characteristic of multicellular, sexually reproducing organisms that continue to reproduce over a substantial part of their life cycle. It is only in such organisms that there is a considerable portion of the life-cycle during which the force of natural selection is declines with age.

The only universal mechanism involved in senescence is that of its evolution.

To conclude, I think the success of evolutionary biologists in understanding major aspects of the ageing process underscores the dynamism and vast scope of current evolutionary biology, a discipline that is fundamentally different from all others in biology because, when faced with any phenomenon, it asks the question WHY?, rather than HOW? I hope that this article has left the reader with some feeling for the impact that evolutionary biology can have on fields not traditionally considered to be within its domain.

Suggested Reading

- P B Medawar. *An Unsolved Problem in Biology*. H K Lewis, London, 1952.
- P M Service, E W Hutchinson and M R Rose. Multiple genetic mechanisms for the evolution of senescence in *Drosophila melanogaster*. *Evolution*. Vol.42. pp 708-716, 1988.
- M R Rose. *Evolutionary Biology of Ageing*. Oxford University Press. New York, 1991.
- M R Rose. Finding the fountain of youth. *Technology Review*. Vol.95(7). pp 64-68, 1992.
- M R Rose and C E Finch (Ed). *Genetics and Evolution of Ageing*. Kluwer Academic. Dordrecht. The Netherlands, 1994.

Address for Correspondence
Amitabh Joshi
Animal Behaviour Unit
Jawaharlal Nehru Centre for
Advanced Scientific Research
Jakkur P.O.
Bangalore 560 064, India
email: visitor@ces.iisc.ernet.in
Fax: (080) 846 2766

Excuses for Not Doing the Math Homework

1. I accidentally divided by zero and my paper burst into flames.
2. I could only get arbitrarily close to my textbook. I couldn't actually reach it.
3. I have the proof, but there isn't room to write it in this margin.
4. I was watching the World Series and got tied up trying to prove that it converged.
5. I have a solar powered calculator and it was cloudy.
6. I locked the paper in my trunk but a four-dimensional dog got in and ate it.

From Internet

Aerogel

The Lightest Solid Known

D Haranath

D Haranath is presently working as a Junior Research Fellow at the Department of Physics, Shivaji University, Kolhapur. His fascination for porous solids and glasses has prompted him to do research on sol-gel processing of silica gels and its derived glasses.

Aerogel, a material not much denser than air on a foggy morning, may provide a good number of applications to the scientific community.

It certainly looks like a whiff of smoke frozen into immobility. If you happen to hold a tile of silica aerogel, commonly called silica air-glass, you would feel practically no weight in your hands. But your mind may well be loaded with a number of questions about its origin, its structure and its properties.

Aerogel is a highly porous solid composed of 0.2% microscopic strands of silicon dioxide as a tenuous web and 99.8% air. This next to nothing solid is due to a gel from which all the liquid has been removed, leaving only a porous framework of silica with air-filled interstices. This has a density as low as 5 milligrams per cubic centimeter; that is only 4–5 times greater than the density of air at sea level. Despite its seeming lack of substance, it is strong enough to support a weight 1600 times its own weight!

Flash Back

As early as in 1864, T Graham had shown that the water in silica gel could readily be replaced by organic liquids. Biologists have taken advantage of this discovery and have successfully replaced the water from gelatinous tissues by alcohol, xylene and paraffin. The final product is a gel in which the organic material is the disperse phase instead of water (see*Box 1* for some definitions).

These facts had led S S Kistler in 1932 to the conviction that a gel, once formed, is independent of the fluid filling its meshes and that fluid might as well be a gas instead of a liquid.

Box 1

A *colloid* is a suspension in which the dispersed phase is so small (1–1000 nm) that gravitational forces are negligible and interactions are dominated by short-range forces, such as van der Waals' attraction and surface charge interactions.

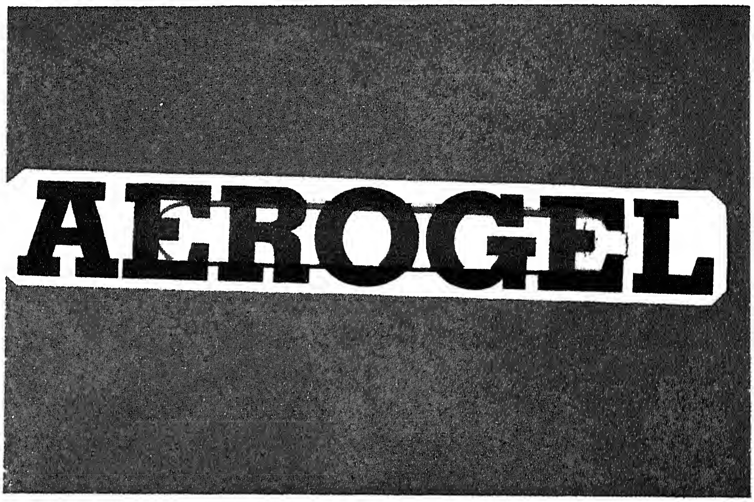
A *sol* is a colloidal suspension of solid particles in a liquid.

An *aerosol* is a colloidal suspension of particles in a gas (the suspension may be called a fog if the particles are liquid and a smoke if they are solid).

A *gel* is difficult to define precisely, although even children intuitively recognise gels. A possible description of a gel is that it is a substance that contains a continuous solid skeleton enclosing a continuous liquid phase. Typically, a polymer molecule which has reached macroscopic dimensions so that it extends throughout the solution is a gel. However, covalent linking to form a giant molecule is not a pre-requisite. Particulate gels held together by van der Waals' forces and gels formed by entangled chains are possible.

This was an important insight. But then came the tricky part; the liquid had to be removed and substituted by a gas without modifying the gel structure. Normally as liquid leaves a gel, surface tension at the liquid-vapour interface causes considerable shrinkage upon drying, making the network collapse on itself. The solution to this problem is to dry the gel at a fairly high temperature and pressure so that the liquid is in a supercritical state. In such a state there is little difference between a liquid and a gas, leading to minimum effect on surface tension. Hence, the molecules of the liquid can be slowly removed from the gel without disturbing the porous network.

Kistler's method of producing aerogels was very tedious and took several weeks to finish. A significant improvement was achieved in 1962 by Teichner and coworkers. Teichner had been approached by the French Government to design a method to store rocket propellants in porous materials. He succeeded in designing a new method to speed up the process of making the gel.



A 10 mm diameter sample of silica aerogel prepared in the author's laboratory. Note the transparency and monolithicity of the gel.

Preparation and Properties

There are several ways in which gels can be prepared. Here we shall focus on a process based on alkoxide synthesis. An alkoxide precursor such as $Si(OR)_4$ where R can be an organic group like methyl, ethyl, propyl, etc., is hydrolysed by mixing with water in the presence of the corresponding alcohol (equation 1).

$$(OR)_3Si–OR + H_2O \longrightarrow (OR)_3Si–OH + ROH \quad (1)$$

$$(OR)_3Si–OR + HO–Si(OR)_3 \longrightarrow (OR)_3Si–O–Si(OR)_3 + ROH \quad (2)$$

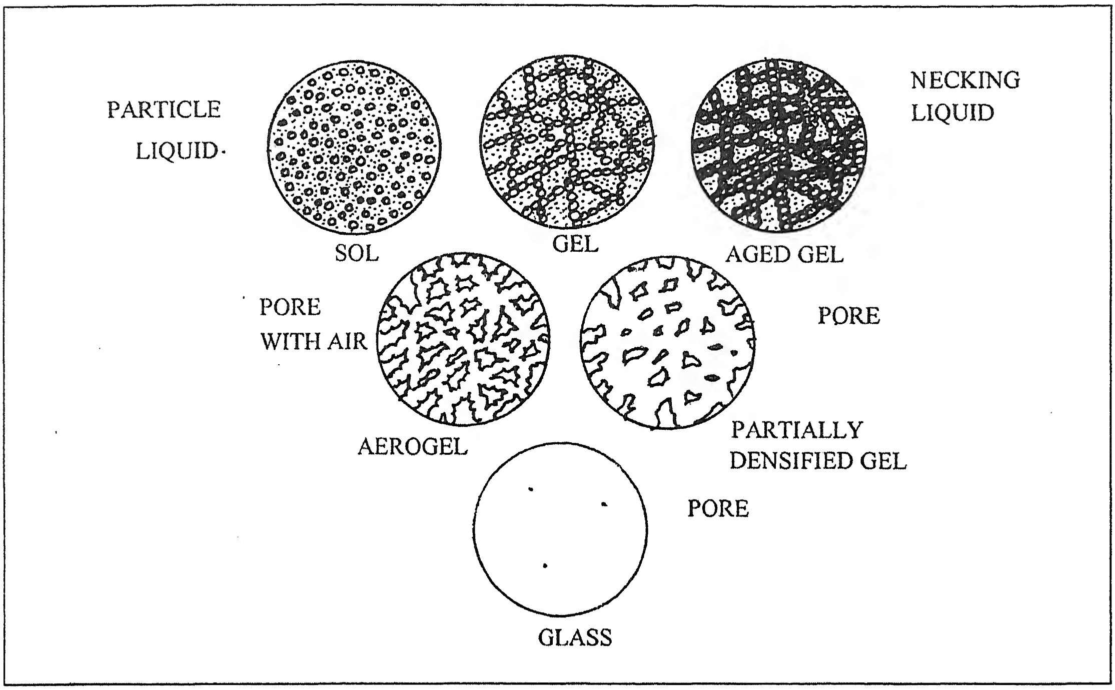
$$(OR)_3Si–OH + HO–Si(OR)_3 \longrightarrow (OR)_3Si–O–Si(OR)_3 + H_2O \quad (3)$$

The hydrated silicate tetrahedra interact in polycondensation reactions (equations 2 and 3) forming ≡Si–O–Si≡ bonds and eventually resulting in a rigid SiO_2 network. This state is called a *gel* (the state that most interests children!). This gel is then dried in an autoclave at supercritical conditions of the respective alcohol. As a result, a highly porous, low density, large surface area *silica aerogel* is produced.

What is so special about the aerogel? The microstructure of the aerogel resembles a bunch of pearl necklaces heaped on a table.

- Aerogel can be used as effective and selective catalysts.
- Aerogel dust in grain and seed stocks was found to kill insects by mere water extraction from the organic tissues.
- For the detection of fast pions, kaons or protons, a medium with a refractive index close to that of air is required. Aerogel exactly fits within this range.
- Aerogel used as core material for windows would make them sound proof and also impermeable to heat.

Figure 1 SOL-GEL processing and resulting structures.



The particle size and pore size are a few nanometers (billionths of a meter). *Figure 1* provides a comparison of the internal structures of a typical sol, different types of gels and glass.

The irregular chain-like structures in aerogel lead to some unusual properties. Aerogels are poor conductors of both heat and sound. The thermal conductivity of a good quality aerogel is nearly 9 mW/mK and the velocity of sound is as low as 100 m/s. It mostly reflects sound. When a piece of aerogel is dropped from a height, it produces a metallic ringing noise. Aerogel is transparent because the size of the structural entities are smaller than the wavelength of visible light. But due to Rayleigh scattering, the substance has a bluish tint. Aerogel has a very low refractive index ranging from 1.01 to 1.08 and a large surface area per unit volume varying from 600 m^2/g to 1000 m^2/g. These are two unusual combinations of properties of a gas and a solid.

Aerogel has a very low refractive index and a large surface area per unit volume. These are two unusual combinations of properties of a gas and a solid.

Applications of Aerogels

Aerogels are being developed as catalysts to promote chemical reactions, as sound insulators, as elements in sonic range- finding devices used by automatic focus cameras, and as Cherenkov radiation detectors in nuclear reactors. Another potentially valuable application of aerogel would be insulation in refrigerators, replacing foam plastic.

Aerogels are being used by NASA for collecting micrometeoroids in space. Because of their low density, aerogels should be able to capture the tiny, fast-moving particles without damaging them. As aerogels are highly transparent, the captured micrometeoroids and their paths through the material can be studied easily.

Aerogels can be densified to ultra high pure glass at relatively low temperatures (1200°C) when compared to commercial glass manufacture (which is done at 2000°C). Partially densified porous glass provides a host matrix for the incorporation of organic or inorganic species, for a variety of applications.

The future potential of this novel aerogel material is endless. Its fascinating features make possible applications, either directly or as a host material, in a wide range of optical products, lenses, wave guides, optical fibers, filters, dye lasers and nonlinear optical devices. One can expect many more new hybrid optical components with multiple functions in the next few years.

A particularly nice feature about silica aerogel is that it is environment-friendly. The substance consists of the same structural entities that make up common beach sand. When exposed to water the material simply disintegrates into fine sand. What better hi-tech material could one ask for?

Suggested Reading

- R K Iler. *The Chemistry of Silica.* Wiley. New York, 1979.
- D W Schaefer and K D Keefer. *Better Ceramics Through Chemistry.* Elsevier-North Holland/ N. Y., 1984.
- J Fricke. Ed. *Aerogels.* Springer Verlag. Heidelberg, 1986.
- C J Brinker and G W Scherer. *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing.* Academic Press. Boston, 1990.

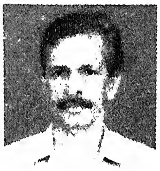
Address for Correspondence
D Haranath
Department of Physics
Shivaji University
Kolhapur 416 004, India

Talipot: A Forgotten Palm of the Western Ghats

A Plea for its Conservation

M D Subash Chandran

A beautiful monocarpic palm of the Western Ghats, talipot which through the ages played a silent role in the culture, economy and ecology especially of the west coast of South India, is facing an uncertain future mainly due to negligence. Much could be done to bring back this palm from obscurity into the mainstream of conservation .



M D Subash Chandran combines the teaching of Botany to undergraduate students with an active interest in vegetational changes and forest history especially of the Western Ghats. He is an associate of the Centre for Ecological Sciences of the Indian Institute of Science, Bangalore.

Introduction

In the outskirts of Palakkad in Kerala, not far away from my home, amidst the rice fields and swaying palmyras, an unusual sight appeared one day – a talipot or tali palm in bloom. A gorgeous, creamy yellow inflorescence, perhaps unmatched in size by any other plant on earth, sprang from the crown of the palm. The pompous display lasted for several weeks; a profusion of tender green tiny fruits followed, gradually reaching the size of ping-pong balls. As months passed the enormous leaves drooped and died one by one. After nurturing the fruits to maturity, the palm itself died!

Years ago, though rare in Palakkad, the tali palm occurred in the hamlets of weavers who made leaf mats and umbrellas. In the evenings of late May, as lightning flashed and thunder roared, heralding the imminence of monsoon rains, people would throng around a bullock cart heaped with leaf umbrellas, making their selections. In the gusty winds of the rainy months, it was not unusual to see school children chasing their leaf umbrellas spinning away on the veranda. Today, except for a dome like version used by the women workers in the rice fields, tali leaf

A gorgeous, creamy yellow inflorescence, perhaps unmatched in size by any other plant on earth, sprang from the crown of the palm.

umbrellas are not found anymore. The tali palm has become rarer too in the otherwise palm fringed skyline of Kerala.

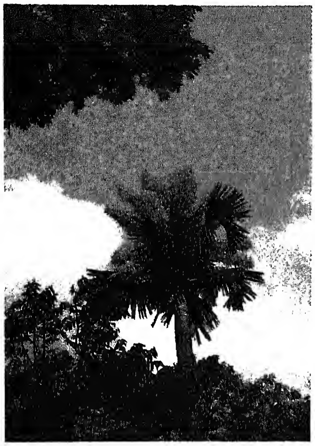


Figure 1 Talipalm, Corypha umbraculifera in flowers.

Years later, I was thrilled to see hundreds of tali palms in Yana, a forested village in the Kumta taluk of Uttara Kannada (North Kanara) district of the central Western Ghats. Prominent in the semi-evergreen hill forests, and clearances, tali palm also occurred in the dwellings of Kumri Marattis, who were once shifting cultivators. Mature palms in bloom could be spotted from kilometers away. I was in the home-range of tali palm which is indigenous to some of the forests of Kumta and Honavar taluks in Uttara Kannada.

Corypha umbraculifera, the talipot or tali palm, mostly planted, occurs along the east coast of India upto West Bengal and also in Sri Lanka and Myanmar. The palm is tali in Bengali, Kannada and Marathi. In Bengali it is also called 'bajarbatur'. It is 'kodapana' in Malayalam and 'kudaipanai' in Tamil; both mean 'umbrella palm'. The Telugu name is 'shritalam'. The tali palm is not to be mistaken for the palmyra, *Borassus flabellifer* which too shares the local name'tali'.

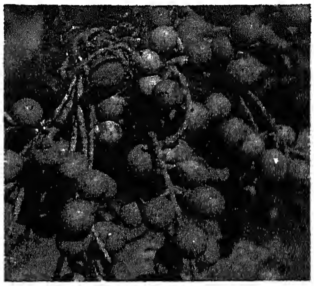
Habitat and Morphology

Figure 2 Talipalm, Vegetative.

The tali palm in Uttara Kannada generally favours the semi-evergreen forests along the spurs and slopes of the Western Ghats, from near the sea level to 600 meters. It occurs both on good soil as well as on the eroded and stony slopes with granite, schists and quartz rather than on exposed laterite. Rainfall in its natural zone in Uttara Kannada is between 3000 to 5000 mm.

The trunk of the palm is 0.6 to 0.9 m in diameter and 15 to 24 m tall. The leafy crown adds 6 to 8 m more to the total height. The fan-shaped leaves have a diameter of 3 to 5 m. They are cleft half way upto the middle into 80–100 linear segments. The specific name 'umbraculifera' means 'open umbrella', obviously an allusion to the leaves. The petiole, about 3 m long and channeled along its upper side, has sharp and saw like margins.

Figure 3 A bunch of young fruits.



The tali palm is monocarpic since it dies after flowering and fruiting. The inflorescence, a pyramidal spadix 3 to 6 m high, springs from the centre of the leafy crown. The several branches of the spadix are covered with millions of minute flowers. The flower has a three toothed calyx and 3 petals, each about 2 mm long. It is bisexual with six free stamens and a gynoecium of three fused carpels. The three chambered ovary has an ovule in each chamber. The ovary narrows into a style which ends in the stigma. The flowering begins with the hot season, although an occasional palm might flower at any time of the year.

Only one of the three carpels matures into the fruit, a drupe 4 cm in diameter. It has a single hard, white, smooth and polished seed with the texture of ivory. The seed is dispersed by birds, bats, squirrels, porcupines, and many other herbivores, which feed on

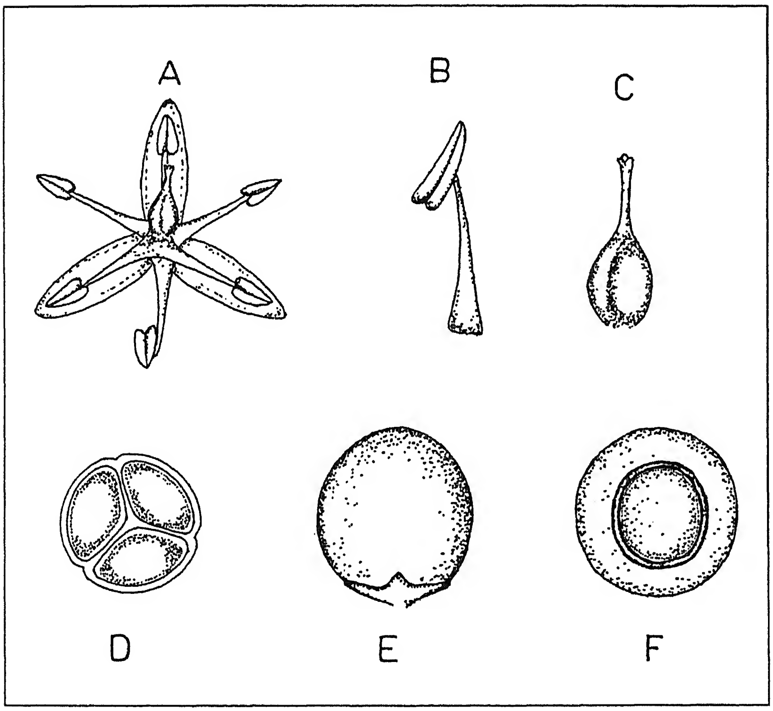


Figure 4 A. Flower; B. Stamen; C. Gynoecium; D. Cross Section of Ovary; E.Fruit; F.Cross Section of Fruit.

the fleshy fruit. The rain water rushing down the steep hill slopes also disperses the seeds.

Economic Importance

The many uses of the tali palm made it popular in the past. The leaves are used for making umbrellas, baskets, mats, fans, coverings for fire-crackers, and for thatching. The leaves were once used for writing. Sacred Buddhist and Hindu texts and ancient medical works were written down on palm leaves which are still to be found in museums, archives, temples and in the households of traditional scholars and medicinal men of South India.

The tali palm is monocarpic since it dies after flowering and fruiting.

In the mature tali palm the stem swells towards the centre storing starch. Over 250 kg of edible starch could be extracted from the pith of a fully grown palm. The light brown starch once formed an important item of food for thousands of people in Uttara Kannada, mainly the forest dwelling Kumri Marattis and many poorer people of the coast. The starch is cooked into a gruel or flattened into bread. It was once used for making 'gulal', a red coloured decorative powder used for ceremonial occasions, for the making of which there was a factory in Honavar. Forming a substitute for ivory, the seeds were once used for making buttons and beads and for miniature carvings. These were also exported from the region. The pounded fruit paste is a fish poison. The handsome palms are good for landscaping.

On the Conservation of Tali Palm

Due to greater availability of food grains, for the last many years, the Uttara Kannada people have hardly cut down any tali palm to extract starch. Whereas the palm is having a new lease of life in its natural habitats of Kumta and Honavar, in other places where it used to be planted, it is almost forgotten. Moreover the strengthening cult of exotic trees has pushed the tali palm into obscurity.

At one time both the British conservationists and the users of tali palm along the west coast, like the umbrella makers and weavers, had planted and protected these trees. In 1878 Colonel Byrde saved several young palms from getting destroyed by a railway line in Sri Lanka and planted them in the market grounds of Kandy. Tali palms were at one time found frequently in the gardens of Kerala. In 1880 some palms were planted by the Sirsi municipality in Uttara Kannada. They used to be present in the gardens of Honavar town. A P Benthall states that in Bengal the palm was more common once and later it vanished except from the Royal Botanic Garden and Eden Garden in Calcutta. In 1942, the last palm outside these gardens, a 12 year old one near Alipore police station, was cut down.

At one time both the British conservationists and the users of tali palm along the west coast, like the umbrella makers and weavers, had planted and protected these trees.

In Uttara Kannada, shifting cultivation was totally forbidden by the close of the last century creating hardships to the poorer peasants like the Kumri Marattis. From the hill tops and slopes, where they grew millets, legumes and vegetables, they had to come down into the valleys to work for others or for settled farming. As a consequence several villages got deserted and fields turned into fallow land. The need for palm starch shot up phenomenally among the forest dwellers as well as poorer people of the coast. To alleviate their miseries certain concessions were allowed. In 1903 the Kumri Marattis of Honavar and Bhatkal were allowed to take one palm each (for the sake of flour) every year free of charge. A total of 1477 Marattis were eligible for this concession. These rules were soon changed allowing one palm to an adult and one-third of a palm to every child under 12 years of age. Excess palms were allowed to the Marattis at Rs.1 per tree and for others at Rs.2 per tree.

About 15,000 palms, yielding an estimated 150,000 headloads of pith, were cut down in the Honavar forests during 1899-1901. At one time palms were seldom cut unless they yielded 16 or more headloads of pith. But soon palms much younger, yielding just 5 to 10 headloads were felled. Mature palms survived only in almost inaccessible places. Awakened by this threat to the tali palm the Government of Bombay designated a forest officer, R S Pearson, to make a plan for conservation and sustainable use of the palms. Pearson commenced his work in 1906; but it seems to have been completed by P E Aitchison in 1908. Their work titled *Working Plan for the Honawar Tali Palm (Corypha umbraculifera) Forests* may be considered as one of the earliest conservation documents from our country. The plan aimed at systematizing the exploitation of the palm so that there would be a regular and unfailing supply to meet not only the wants of the local people but to ultimately bring the forests into their normal condition.

Due to the strict regulations on the exploitation of tali palm and the greater availability and production of food grains, the palms are not cut down for food anymore.

Pearson classified the tali palms of Honavar under four classes: Class I - palms containing at least 8 head loads of pith (18,559 palms); Class II - full-grown palms with less than 8 head-loads of pith (58,230 palms); Class III - half-grown palms (91,615 palms) and Class IV - young palms past the seedling stage (184,113 palms). The plan prescribed exploitation of only the 'mature' palms. In the Kandy market grounds of Sri Lanka some of the palms were reported to have flowered after 38 years while others did not. Pearson initiated an experiment to estimate the age of the palms. The trunks of the palms are covered with petiole scars. If the number of leaves produced by a palm in a single year was known one could probably arrive at their approximate age by counting the total number of leaf scars. The study was continued by Butterworth, who reported in 1915 that the flowering age would be about 88 years. Opinions still differ about the flowering age of tali palm, which is put between 40 and 90 years.

Due to the strict regulations on the exploitation of tali palm and the greater availability and production of food grains the people in the palm belt of Uttara Kannada do not cut down the palms for food anymore. The tali palm forests appear to be well stocked. Yet it should be noted that the palm was once associated with the shifting cultivation areas. Since the last one hundred years there is no slashing and burning of forests in the palm belt and the evergreen forests are on the return. It is not clear what is going to be the fate of the tali palm. What are the other threats to the palms today? Population studies need to be urgently carried out on the palm in its natural areas.

The potential area for its cultivation is vast consisting of the humid parts of the Indian peninsula, and perhaps the Assam region. The tree with its very useful leaves and the stem stocked with starch holds great promise to sustain native arts and crafts and supplement nutrition in the tribal areas of tropical India. The very presence of mature tali palms could be a reassuring sight in the famine prone areas. If the starch is not extracted before flowering the large output of fruits (over 200 kg per palm) offers food to a variety of wild life such as birds, bats, porcupines, squirrels, boars, deers and sambar. Sadly tali palm as an ecological resource is almost forgotten in the forestry circles. Since the fruits which take nearly a year to ripen are consumed at all stages by a variety of animals the tali palm has also the potential to be developed into a keystone resource in the tropical forest belt of India.

The tree with its very useful leaves and the stem stocked with starch holds great promise to sustain native arts and crafts and supplement nutrition in the tribal areas of tropical India.

Suggested Reading

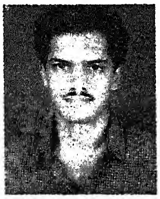
- R S Pearson and P E Aitchison. *Working plan for the Honawar Tali Palm (Corypha umbraculifera) forests*. Forest Department of the Bombay Presidency, 1908.
- W A Talbot. *Flora of the Bombay Presidency and Sind*. Vol. 2. Government Photozincographic Press. Poona, 1909.
- E Blatter. *The Palms of British India and Ceylon*. 1926. Repr.1978.
- J Jenik and M D S Chandran. Plan for talipot palm. *Threatened Plants Newsletter*. No.19, IUCN, 1988.

Address for Correspondence
M D Subash Chandran
Department of Botany
Dr Baliga College of Arts and Science
Kumta 581 343, Karnataka, India

What's New in Computers

Flash Memories

Vijnan Shastri



Vijnan Shastri works at the Centre for Electronics Design Technology in the Indian Institute of Science, Bangalore, and his areas of interest are multimedia systems, microprocessor systems, storage subsystems and systems software.

Flash memories are not very new. Products based on flash memory have been around for more than 4 years now after the basic technology was introduced in 1988. Although flash-memories were initially thought of as replacements for magnetic storage, this has not happened. Instead, what is new are the different applications in which flash memory is being putting to use. What is flash memory? What are its features ? What are its advantages as compared to conventional memories? Where is it used ? This is what we'll take a look at in this article. The actual internal structure of flash memory is illustrated at the end of the article.

Magnetic Storage, RAM and EPROM

Disk memory and RAM memory are two types of read/write memories which we are all familiar with. We also know that this disk memory is non-volatile i.e., the data stored on the disk does not vanish when the power goes off (due to properties of the magnetic medium). The disk memory consists of several platters of magnetic medium rotating at a certain speed and a read/write head moves across the disk, reading and writing data. Currently, the size of such magnetic disks is in the range of 1–2 Gigabytes for PCs. RAMs are solid-state Integrated Circuits (chips), much faster than magnetic disks and are typically installed in sizes of 8–16 Mega bytes on desktop PCs. These RAMs are however, volatile. There is a type of commonly used non-volatile, solid-state memory called EPROM (Erasable Programmable Read Only Memory). But EPROMs need to be erased (using UV-light erasers) in order to be written into. They cannot be written into *in-situ*. Their structure is similar to that of flash memory cells except that the oxide layers are thicker than those in flash memory cells. Various types of memories and their main features are summarized in the *Box*.

Types of Read /Write Memories

Magnetic Disks

- Typical capacity is 1-2 Giga bytes of storage (for PCs), Rs 9 per Mbyte, access time is in the order of 20 ms.
- Not very rugged due to moving precision mechanical components.
- Typical read rate is 5 Mbytes/s, write rate is about 3–4 Mbytes/s.
- Power consumption is about 3 W.

RAM (Random Access Memory) is of two types: DRAM and SRAM

- *DRAM*
  - Dynamic RAM, High Density (1 transistor per memory cell), access time is 40–70 ns, Rs 300 per Mbyte. PCs are now equipped with 16–32 Mbytes of DRAM.
  - Disadvantage is that it must be *refreshed* periodically. This implies that it is not available for access by the processor for read/write cycles during refresh. DRAM is volatile (no-power implies memory gets erased).
- *SRAM*
  - Static RAM, Low density (6 transistors per memory cell), used in low-capacity applications such as buffers, caches and FIFOs (Typical capacities: 256–512 kbytes). It is volatile.
  - Access time is 10–100 ns (depending on type) and does not need refresh cycles.

EPROM (Erasable Programmable Memory)

- Has to be erased (by UV light) and can be written into using special circuitry. It is non-volatile. Typical capacities are 32–128 kbytes.

Flash Memory

- Solid-state (and hence rugged) high density (one transistor per memory cell) relatively low-power (350 mW–max).
- Rs 700 per Mbyte.
- Typical capacities are 2 Mbyte (for digital cameras) to 80 Mbytes (PCMCIA card) for portable computers. PCMCIA card is the size of a credit card and is less than 3.5 mm thick.
- Read rate is 5 Mbytes/s and write rate is 700 kbytes/s.

Flash Memory — Solid-state Yet Non-volatile

Flash memories are solid-state chips but are non-volatile. These properties immediately made designers think of them as substitutes for magnetic disks in portable laptop and notebook computers. Flash memories are built from transistors (using one transistor per memory cell) — the same basic element of RAM chips. But they differ in their structure and operation from RAM to give them their above mentioned characteristic.

Designers produced flash memory PCMCIA (Personal Computer Memory Card International Association) cards, the size of a credit card, in the range of 20 MB–80 MB for portable computers hoping that one day such cards would replace magnetic disks in these machines. However this has not happened and magnetic disks are still being used in portable PCs.

The characteristics of flash memory and magnetic disks are given in the *Box*. Flash memories offer advantages of ruggedness, high density (one transistor per memory cell or bit) and low-power consumption due to their solid-state structure as opposed to magnetic disks which are built from several complex, moving mechanical components. However the storage capacities of magnetic disks have been doubling every year for the past 4–5 years (and continue to do so) for the same volume of storage. They offer an unbeatable price of about Rs 9 per MB, in terms of storage costs as against flash memories with Rs 700 per MB price tag. Further, flash memories have a disadvantage that a limited number of writes are possible as explained in the next paragraph.

Flash memories have the disadvantage that the number of write operations are limited to about 100,000. This meant that designers had to build in lots of spare cells which would increase the life of these memories. They took advantage of the fact that applications do a lot more reads than writes. When a memory cell reaches its write-limit, a spare cell is allotted to take its place. The write to a flash memory is a complex operation requiring voltages of 10 V (above the normal 5 V) for a specified duration. All these make controllers for such flash memories complex and flash memories themselves expensive.

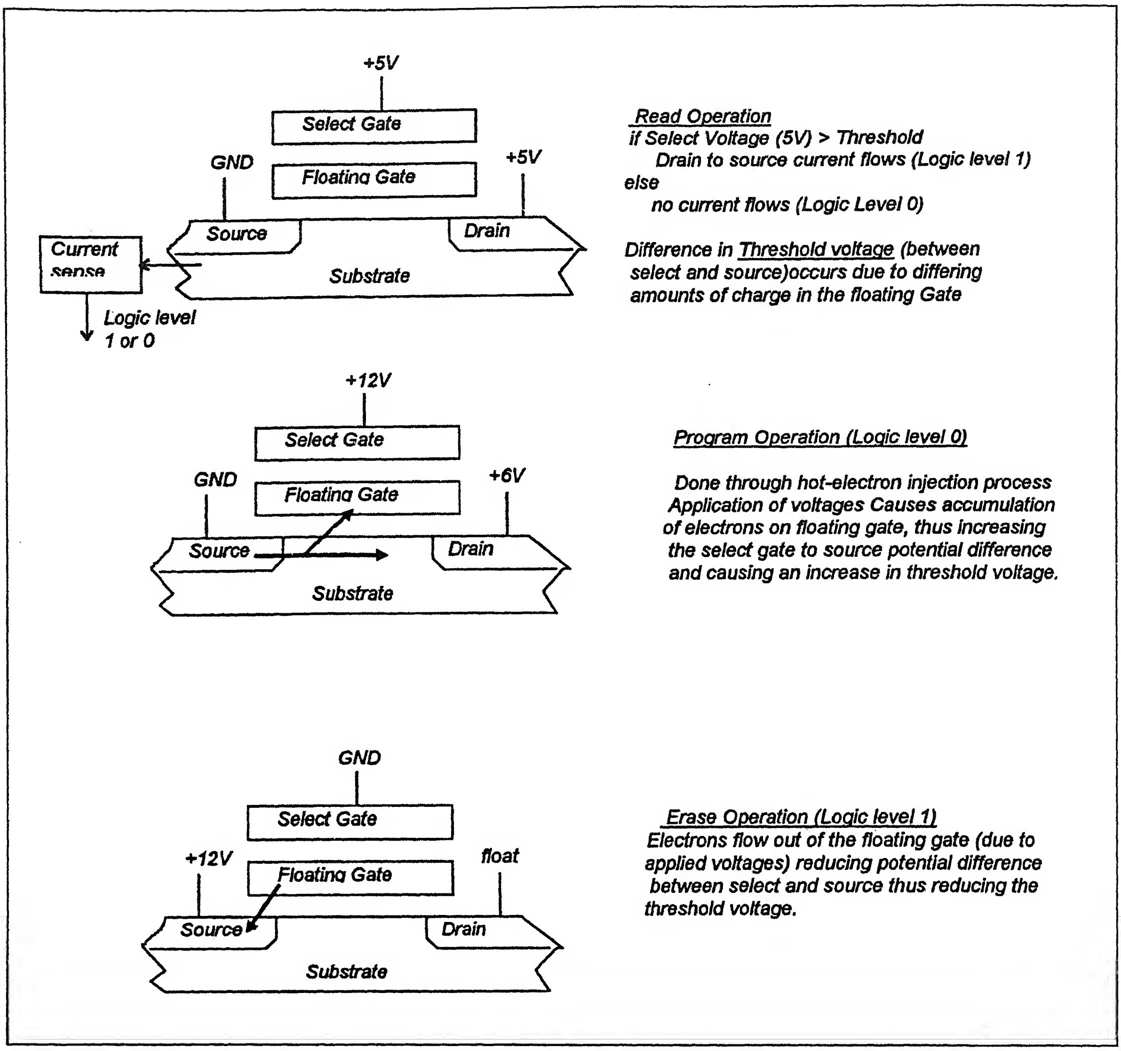
Thus flash memories never took off (as their protagonists hoped) in the storage scenario for portable machines. But they are being used in several other applications– especially in digital cameras, which are expected to be the launching pads for flash memory — as we'll see in the following paragraphs.

How Do They Work?

Flash memory transistors have an extra*floating gate* in addition to the gate that is present in normal transistors (see *Figure*). The floating gate is the charge-storage element which gives flash memory its non-volatile characteristic. However, charging or discharging the floating gate is slightly complex and is illustrated in the *figure*. The complex charging process results in the abnormally long write cycles of flash memory. Hence the write rate is very much less (700 k bytes/s) than the read rate (5 M bytes/s).

EEPROM

A slight variant of flash memory technology has been used to make highly successful EEPROM chips. EEPROM stands for Electrically Erasable Programmable ROM. These chips are replacing all applications where traditionally EPROMs are used. EPROMS are generally used to store firmware. Firmware (also called BIOS – Basic Input Output System – in a PC) is the name given to software which normally is executed first in the PC, when it is powered on. Firmware initializes all the peripheral devices of the PC such as disk, display and the keyboard and enables further loading of the operating system from disk to RAM to take place. Products such as fax machines, laser printers, telephone answering machines, cellular phones, digital exchanges and electronic cash registers require a microprocessor (or



Read Operation
if Select Voltage (5V) > Threshold
    Drain to source current flows (Logic level 1)
else
    no current flows (Logic Level 0)

Difference in Threshold voltage (between select and source)occurs due to differing amounts of charge in the floating Gate

Program Operation (Logic level 0)

Done through hot-electron injection process Application of voltages Causes accumulation of electrons on floating gate, thus increasing the select gate to source potential difference and causing an increase in threshold voltage.

Erase Operation (Logic level 1)
Electrons flow out of the floating gate (due to applied voltages) reducing potential difference between select and source thus reducing the threshold voltage.

microcontroller) to manage the machine. The programs for these controllers (while certainly not as complex and large-sized as PC operating systems such as DOS or Windows 95) do require storage, and traditionally a combination of EPROM and battery-backed up RAM are used to meet the demands of these programs. The latter type was used for data that needs to be changed from time to time but needs to be non-volatile. EEPROM is now replacing this combination of memories since it offers non-volatility of storage. A further advantage of using EEPROMs is that if there are bugs in a program they can be corrected without removing the chip from the machine unlike EPROMs which have to be removed from the system in order to be re-programmed.

Digital Cameras

Recently, flash memories are being used in digital still-picture cameras – replacing films. In appearance, these digital cameras look exactly like the normal camera. They acquire the image through their CCD (Charge Coupled Device) cell array and transfer the image to a flash-memory card – which serves as the *film*. The image can then be viewed on a TV or transferred to a PC for archiving into electronic *photo albums*. Some of the more sophisticated models even have a LCD screen to view the image instantly after the photograph is taken. The photographer can review his work 'on-line' (and do a re-take if necessary). In their digital form these images can be easily processed using image processing techniques for either efficient storage or for feature extraction. The implications of these are enormous. This has a wide variety of applications in fields such as medical databases, advertising, police crime records, voter identity card preparation and verification etc. This is especially true if this technology is combined with simple magnetic card storage technology which has even been used for producing bus and train tickets. In future, (when prices permit) low-capacity flash-memory cards may also be used in place of these magnetic cards. Another major advantage of these digital images is that they can be easily transmitted over modem or the internet to a remote location without distortion or delay. This is especially useful to journalists who just need a simple portable PC with a modem to transmit images to the main office. Thus the possibilities are many, the applications are exciting and flash memory will have a powerful impact in these areas. This explains why all the major photographic companies which have concentrated on film-based photography for decades, have now moved in a major way to the digital camera arena.

The negative point is that these digital cameras are still very expensive. A model recently released in India, which can store about 100 images costs about Rs. 22,000. This is still prohibitively expensive. However, prices are expected to reduce with about 20 vendors entering the market — including the major photo companies.

In this article we have focused on the new applications of a not-so-new device – flash-memory. Flash memory is still searching for the application that will increase its demand to high levels in the mass market. Many believe that the digital camera is that 'killer' application. Manufacturers of flash memory hope that the expected high demand in the consumer market for flash memories will bring down its current high price and will enable it to compete in the PC market as well. They even predict that with improvements in technology that will overcome its current limitations, flash memory could even replace DRAM (Dynamic-RAM) in computers. This is because flash memory cells have the same density as DRAM cells and would occupy the same volume as DRAM. Time, and the response of the consumer will tell if flash memory will really make the big splash it's protagonists want it to make.

Address for Correspondence
Vijnan Shastri
Centre for Electronics
Design Technology
Indian Institute of Science
Bangalore 560 012, India
email:
vshastri@cedt.iisc.ernet.in
Fax:(080) 334 1683

Suggested Reading

- Brian Dipert and Lou Herbert. *Flash memory goes mainstream.* IEEE Spectrum, October 1993.
- Gary Legg. *Flash Memory Challenges Disk Drives.* EDN February 1993.
- Rick Cook. *The way of all flash.* Byte Magazine, June 1996.

Science is facts; just as houses are made of stones, so is science made. of facts; but a pile of stones is not a house and a collection of facts is not necessarily science. *Henri Poincare*

The most exciting phrase to hear in science, the one that heralds new discoveries, is not Eureka! (I found it!) but That's funny...

Isaac Asimov

Molecule of the Month

Cryptatium, the First "Elementoid"

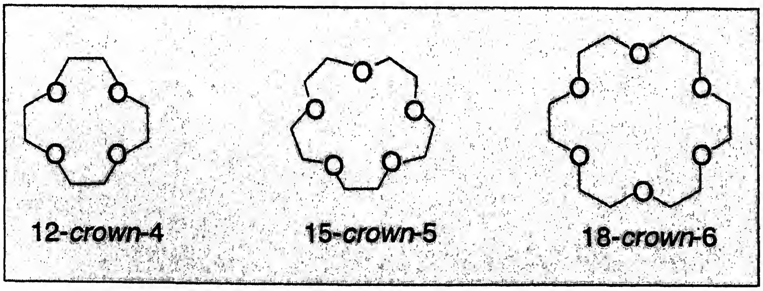
Uday Maitra

Electrochemical reduction of a cryptate yielded a neutral species with the properties of an *expanded* atom.

Uday Maitra is at the Department of Organic Chemistry, Indian Institute of Science, Bangalore 560 012, India

Most of the 92 elements, excluding the transuranium elements, were discovered by the turn of the nineteenth century. The early part of the twentieth century witnessed the discovery of the transuranium elements, but most of them were largely of academic interest because of their high radioactivity, and relatively short half-life. Thus, the search for the heaviest element (or metal?) became rather difficult to pursue.

As a matter unrelated to the discovery of elements, towards the end of the sixties, a rather novel metal ion complexing agent was serendipitously discovered by Charles Pedersen. He found that cyclic polyethers were able to complex alkali metal ions, by providing a cavity in which the charged alkali metal cation can sit comfortably in a *sea* of electrons donated by the oxygen atoms! In a sense this is similar to the hydration of a metal ion in water, except that the stabilization is done by oxygen atoms made available in a cyclic molecule. These crown ethers show high selectivity in complexing alkali metals, since the size of the cavity (which is determined by how big the ring is, it can be 12, 15, 18, 21.. membered) can be made just optimum to match the

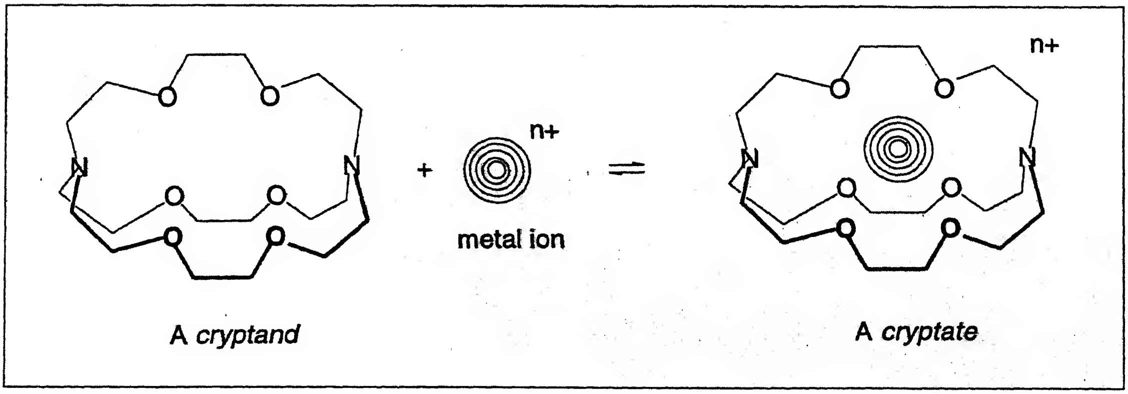


Cryptates can be regarded as *expanded* atoms, where the positive charge of the metal ion has spread over many atoms.

ionic radius of the alkali metal ion. Thus, 12-crown-4, 15-crown-5 and 18-crown-6 have been found to be selective for binding Li^+, Na^+ and K^+, respectively. The ability of the crown ethers to complex alkali metal ions selectively has led to many applications – one of the most important being the solubilization of inorganic salts in organic solvents in the presence of crown ethers. One can, for example, dissolve $KMnO_4$ in benzene in the presence of a crown ether to give a purple solution!

Just when Pedersen was investigating crown ethers, Jean-Marie Lehn was busy creating a three dimensional *cage* for metal ions. His idea was to construct a *cryptand* (Latin *crypta* means cavity; and Greek *kryptos* means hidden) for efficient complexation of alkali metal ions. If a crown ether could be considered as a disk like object, the cryptand is a spherical object, which should be capable of capturing a metal ion more effectively. Indeed, Lehn had found that cryptands formed *cryptates*, alkali metal ion complexes of exceptional stability.

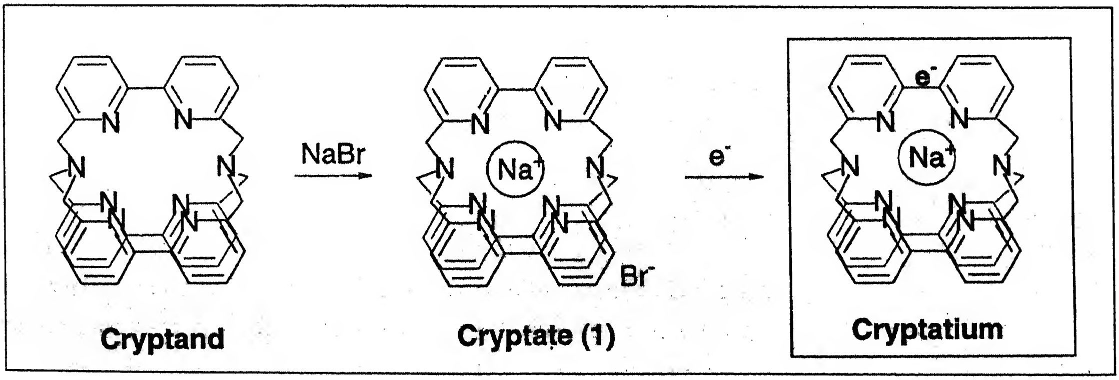
Cryptates can be regarded as *expanded* atoms, where the postive charge of the metal ion has spread over many atoms. One interesting property of such metal complexes is that they interact very weakly with anions or with solvent molecules, since the metal ion is well stabilized inside the cage having a number of oxygen and nitrogen atoms which can donate electron pairs to it! As a result, such (almost) spherical metal complexes can be



regarded as *super-heavy* alkali-metal ions (compare the ionic radius of Cs^+ (1.65Å) with that of the cryptate shown in the previous page which is about 5Å)! If that is so, then it is interesting to think what would happen if an electron is added to a cryptate. Where will it go? Such systems have actually been made by Lehn, and are considered to be salt-like species with the electron as the anion (an *electride*). Such a species can also be thought of as a large alkali metal with a very small ionization potential (more like the Rydberg state of an atom). We therefore see two extremes: a true metal atom where the valence electron is in a metal orbital, or an electride (electron added to a cryptate) where the electron is at a very large distance from the metal cation. Lehn has considered another possibility in between, in which the *valence* electron will be associated with a ligand orbital. The cryptand shown above does not have a vacant low energy orbital for adding an electron, but one can certainly design another cryptand capable of accommodating an electron on its surface..

The addition of an electron to a cryptate can produce a species behaving like the Rydberg state of an atom.

The cryptand shown below essentially has three bipyridine (bpy) units arranged as the three blades of a propeller, with the two ends of each of the three units being linked to a pair of nitrogen atoms. Bipyridyl units have been used extensively for the complexation of metal ions (many of you have probably seen the blood-red color of the bpy complex of ferrous ion, $Fe(bpy)_3^{2+}$).



The bipyridine unit also has a relatively low-energy vacant orbital in which an electron can be added. How does one add an electron to a cation? This is nothing but a reduction process. Normal reducing agents can often create problems, because of the byproducts formed from them. A method which is commonly used for the generation of reactive metals is electrochemical reduction (readers may recall that reactive metals such as Na, K, Ca etc. are always manufactured this way). Indeed, this was the method employed by Lehn and coworkers for adding an electron to cryptate **1**. Upon electrochemical reduction of **1**, they observed the formation of air-sensitive blue-violet crystals on the cathode. Using special techniques necessary for the handling of air-sensitive solids, they were able to get an X-ray crystal structure of this blue-violet crystal. Unlike the crystal structure of a related cryptate bromide salt, the crystals of this blue-violet solid did not show any anion in between the cryptate units – suggesting that this species is electroneutral. This would be possible if the negative charge,*i.e.*, the electron is residing in the molecule itself, and naturally the most likely site for it is the ligand orbital. Support for this hypothesis comes from the observation that in the crystal structure, the Na^+ ion is closer to one bipyridine unit. This is because the bipyridine unit with the extra electron draws the counter ion closer to it. Lehn and coworkers termed this compound *sodio-cryptatium*, and have suggested that this species may be described as an expanded atom, or as a radical contact ion pair. The cryptatium therefore represents a *neutral* species somewhere in between an atom, and an electride. It can also be regarded as a *molecular element*. Lehn has proposed that it might also be possible to extend the list of such elements by incorporating a doubly or a triply charged cation in the cryptand, and reducing the cryptate by adding two or three electrons, respectively.

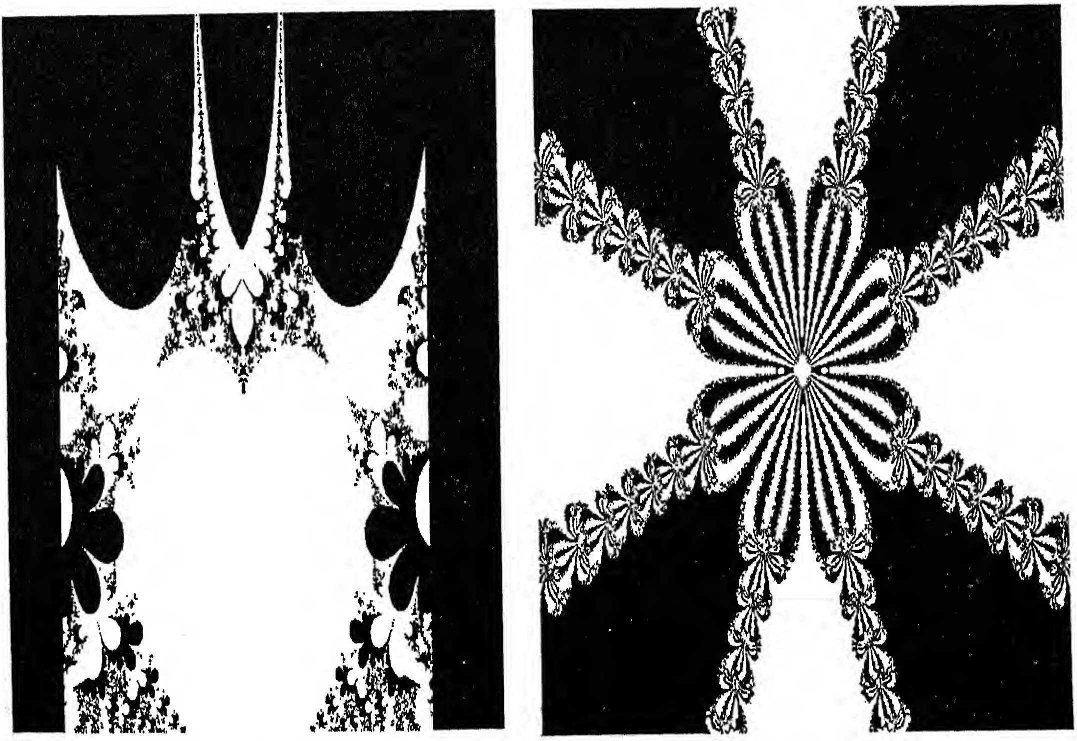
The cryptatium represents a neutral species somewhere in between an atom, and an electride. It can also be regarded as a molecular element.

It is needless to say that such molecules are likely to be of great interest for studying their magnetic, electrical and other properties. Detailed studies should help us understand how exactly the electron gets delocalized over the three bpy units. Lehn has also envisaged a *fullerium* species, which can be considered to be a fullerene, such as C_{60}, containing a metal ion M^{n+} in its internal cavity, and n electrons delocalized on the carbon framework!

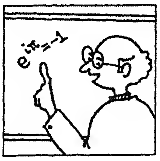
We can perhaps conclude that the search for the heaviest (molecular) element is still on!

Address for Correspondence
Uday Maitra
Department of Organic
Chemistry
Indian Institute of Science
Bangalore 560 012, India
email:
maitra@orgchem.iisc.ernet.in
Fax: (080) 334 1683/309 2690

Fractals



Classroom



In this section of Resonance, *we invite readers to pose questions likely to be raised in a classroom situation. We may suggest strategies for dealing with them, or invite responses, or both. "Classroom" is equally a forum for raising broader issues and sharing personal experiences and viewpoints on matters related to teaching and learning science.*

B Bagchi, Indian Statistical Institute, Banglaore 560 059.

! Bachet's Problem

A grocery shopkeeper keeps five stones of different weights. He is able to use a common balance and weigh out quantities ranging from 1 to 100 kg, in steps of 1 kg. What are the weights of these five stones?

The above is the problem *100 kg with five stones* posed by R Yusufzai in the *Think it Over* column of the July 1996 issue of *Resonance*. A much better problem will result if the figure 100 is replaced by 121. This is because the question *what are the weights of these five stones*? seems to suggest that there are uniquely determined weights to be found! However, as may easily be verified, the weights in kg of the stones might be 1,3, 9, 27 and m, where m is any integer in the range $60 \leq m \leq 81$. In fact, there are many other solutions to the problem as posed. If, however, it was given that the grocer can weigh any object of weight between 1 kg and 121 kg (in steps of 1 kg) using his five stones, then the weights (in kg) of the stones must have been 1, 3, 9, 27 and 81. This is the case $k = 5$ of the result stated and proved below.

The problem is a well-known variation of an old problem due to Bachet (see Suggested Reading). In the original *binary* version, the grocer cannot subtract, so he must put the stones in one pan and the object in the other. Mr Yusufzai's problem is an instance of the *ternary* version where this restriction is removed. The general problem (in its *ternary* version) may be stated as follows: Given a positive integer k, find the largest integer N_k such that any object whose weight is an integer between 1 and N_k (ends included) can be weighed using k stones of suitable integral weights. In this notation, the *Think it Over* problem is to show that $N_5 \geq 100$.

In fact, we have —

Theorem : $N_k = \frac{3^k-1}{2}$. If k stones are such that all integral weights between 1 and N_k can be measured using them, then the weights of these stones must be $3^j, 0 \leq j \leq k-1$.

This is, essentially, Theorem 141 in the book by Hardy and Wright (see Suggested Reading).

In order to prove this, we must convert it into a precise mathematical statement. To this end, let $a_0, \cdots, a_{k-1}$ be the (positive integral) weights of k stones. In order to weigh an object of integral weight m, the grocer places the object together with some of the stones on the right pan (say) and puts some other stones on the left pan. For $0 \leq j \leq k-1$, put $\varepsilon_j = 1$ if the stone of weight a_j is placed on the left pan, $\varepsilon_j = -1$ if it is on the right pan, $\varepsilon_j = 0$ if it is not used. Since the two pans must balance, we get

$$m = \sum_{j=0}^{k-1} \varepsilon_j a_j \quad \text{where } \varepsilon_j \in \{0,1,-1\} \text{ for } 0 \leq j \leq k-1 \qquad (1)$$

This leads us to:

Definition : If $A = \{a_0, \cdots, a_{k-1}\}$ is a finite set of positive integers then the *capacity* $C(A)$ of A is the largest integer M such that for every integer m in the range $1 \leq m \leq$ M, the equation (1) has a solution.

Informally, the capacity $C(A)$ is the largest M such that all weights between 1 and M can be measured using k stones whose weights are in A. In terms of this definition, the above theorem may be restated as follows.

Theorem : If A is of size k then $C(A) \le \frac{3^k - 1}{2}$. Equality holds here if and only if $A = \{3^j : 0 \le j \le k-1\}$.

To prove the theorem, note that if m can be written as in (1) then so can $-m$ (just change the signs of all ε_j); also, trivially, $m = 0$ can be written thus (take $\varepsilon_j = 0$ for all j). Therefore, if $C(A) = M$, then all the $2M+1$ integers m in the range $-M \le m \le M$ can be expressed as in (1). But there are 3 choices for ε_j for each j, hence only 3^k choices for the right hand side of (1). Hence $2M+1 \le 3^k$, or $C(A) \le \frac{3^k-1}{2}$. Now, if we take $A = \{3^j : 0 \le j \le k-1\}$, then

for $1 \le m \le \frac{3^k-1}{2}$ write $\frac{3^k-1}{2} - m$ in base 3: $\frac{3^k-1}{2} - m = \sum_{j=0}^{k-1} \delta_j\, 3^j$,

where $\delta_j \in \{0,1,2\}$. Put $\varepsilon_j = 1 - \delta_j$. Then (1) holds. Thus $C(A) \ge \frac{3^k-1}{2}$ for this set. Together with the previous inequality, we get $C(A) = \frac{3^k-1}{2}$.

Only the uniqueness part of the theorem remains to be proved. In fact, this is the only non-trivial and interesting part. To prove this, let $A = \{a_0, \cdots, a_{k-1}\}$ have capacity N_k. Since, now, equality holds in the inequality $C(A) \le \frac{3^k-1}{2}$ which appears in the statement of the theorem, the proof of the inequality shows that every integer m in the range $-\frac{3^k-1}{2} \le m \le \frac{3^k-1}{2}$ has a *unique* representation (1); conversely any m of the form (1) belongs to this range. Therefore, letting X be an indeterminate, we get

$$\prod_{j=0}^{k-1} (X^{-a_j} + 1 + X^{a_j}) = \sum_{|m| \le \frac{3^k-1}{2}} X^m \qquad (2)$$

as may be verified by multiplying out the left hand. Since, in particular, the largest integer (viz. $\sum_{j=0}^{k-1} a_j$) of the form (1) must be the largest integer in the range $[-\frac{3^k-1}{2}, \frac{3^k-1}{2}]$, we also have:

$$\sum_{j=0}^{k-1} a_j = \frac{3^k-1}{2}. \tag{3}$$

Using (3) and a little algebra, (2) simplifies to

$$\prod_{j=0}^{k-1} \frac{X^{3a_j}-1}{X^{a_j}-1} = \frac{X^{3^k}-1}{X-1}. \tag{4}$$

Now fix $j, 0 \leq j \leq k-1$. Let w be a primitive $3a_j$-*th* root of unity. That is, w is a complex number such that $w^l = 1$ if and only if l is an integral multiple of $3a_j$. (For instance, we may take $w = \exp(2\pi\sqrt{-1}/3a_j))$. Then w is a zero of the left hand side, and hence also of the right hand, of (4). Thus $w^{3^k} = 1$. So $3a_j$ divides 3^k. That is, $a_j \in \{3^i, 0 \leq i \leq k-1\}$. Since this holds for all j, we have $A \subseteq \{3^i : 0 \leq i \leq k-1\}$. Since both sets have size k, we must have $A = \{3^i : 0 \leq i \leq k-1\}$. This proves the uniqueness of the set of given size and maximum capacity.

The reader may like to look up the proof in the book by Hardy and Wright, which is very different from the proof given here. It is a clever use of mathematical induction.

Tail-piece : Bachet is better remembered by mathematicians for another reason. It was on Bachet's edition of Diophantus' Arithmetic that Fermat scribbled his famous marginal notes. Bachet was also the first man to state, (without proof) what is now known as Lagrange's four square theorem: every natural number is the sum of at most four perfect squares.

Suggested Reading

- F Schuh. *The Master Book of Mathematical Recreations*.Dover. New York. pp115-118, 1968.
- G H Hardy and E M Wright.*An Introduction to the Theory of Numbers*. Oxford Univ. Press.London.pp115-117,1971.

Think It Over

This section of Resonance *is meant to raise thought-provoking, interesting, or just plain brain-teasing questions every month, and discuss answers a few months later. Readers are welcome to send in suggestions for such questions, solutions to questions already posed, comments on the solutions discussed in the journal, etc. to* Resonance *Indian Academy of Sciences, Bangalore 560 080, with "Think It Over" written on the cover or card to help us sort the correspondence. Due to limitations of space, it may not be possible to use all the material received. However, the coordinators of this section (currently A Sitaram and R Nityananda) will try and select items which best illustrate various ideas and concepts, for inclusion in this section.*

1 Problem of the Vacillating Mathematician

Question raised in the *Think It Over* section of *Resonance*, Vol.1, No.6.

We thank the readers for their enthusiastic response to this problem. We are not publishing the solution to this problem right now as the solution will be part of an article by K B Athreya on a related theme and will appear in *Resonance* shortly. A correct solution has been provided by Balraj Singh.

2 100kg with Five Stones

Question raised in the *Think It Over* section of *Resonance*, Vol.1, No.7.

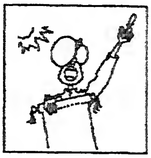
A more general version of this problem has been discussed by B Bagchi in his article *Bachet's Problem* (*Classroom*, this issue). We have received correct solutions from the following persons:
S C Dutta Roy, S D Joshi, Ritesh Kumar Singh, K C Pradeep and G S Kalburgi.

3 Finding the Odd Ball

Question raised in the *Think It Over* section of *Resonance*, Vol.1, No.7.

Discussion on this problem will be taken up in another issue of *Resonance*.

Information and Announcements



National Seminar for Science Writers

Marathi Vidnyan Parishad, Mumbai is organising a National Seminar for Science Writers on December 21 and 22, 1996 in Mumbai. The Seminar will have sessions on various subjects like Science Fiction, Art of Translation, Historical Review of Science Writings and also conversation with science writers. For further details please contact Hon. Secretary, Marathi Vidnyan Parishad, Vidnyan Bhavan, V N Purav Marg, Sion-Chunabhatti, Mumbai 400 022.
(Telephone: 522 47 14, 522 62 68).

1996 Nobel Prizes

Physics

The 1996 Nobel prize in Physics has been awarded jointly to *David M Lee* of Cornell University, USA, *Douglas D Osheroff* of Stanford University, USA and *Robert C Richardson* of Cornell University, USA for their discovery of superfluidity in helium-3.

Chemistry

The 1996 Nobel prize in Chemistry has been awarded jointly to *Robert F Curl Jr.*, Rice University, USA, *Sir Harold W Kroto*, University of Sussex, UK and *Richard E Smalley*, Rice University, USA for their discovery of fullerenes.

Physiology or Medicine

The 1996 Nobel prize in Physiology or Medicine has been awarded jointly to *Peter C Doherty* and *Rolf M Zinkernagel* for their discoveries concerning the specificity of the cell mediated immune defence.

Acknowledgements

Resonance gratefully acknowledges the help received from the following individuals:
Jayant Rao, K Kasturi Rangan, R Mahesh, K V L Sarma and TN Ruckmongathan.

Books Received

The Chemistry of Conscious States
J Allan Hobson
Back Bay Books (Little Brown)
1994, $13.95.

The Arrow of Time
Peter Coveney and
Roger Highfield
Flamingo, an imprint of Harper Col.
1991, Rs.252.

Elemental Mind
Nick Herbert
Blume Penguin
1994, $9.50.

Concepts in Biotechnology
D Balasubramanian, C F A Bryce,
K Dharmalingam, J Green and
Kunthala Jayaraman
Universities Press
1996, Rs.295.

Deterministic Chaos, Complex Chance Out of Simple Necessity
N Kumar
Universities Press
1996, Rs.70.

Mineral Resources of Karnataka
B P Radhakrishna
Geological Survey of India
1996, Rs.200.

Edited and published by V K Gaur for the Indian Academy of Sciences,
Bangalore 560 080. Printed at Thomson Press (I) Ltd., Faridabad 121 007.

Annual Subscription Rates

	Personal	Institutional
India	Rs 100	Rs 200
Third World Countries	US $ 25	US $ 50
Other Countries	US $ 50 (Students $ 25)	US $ 100

Send your subscriptions by DD or MO in favour of
"Indian Academy of Sciences" to Circulation Department, Indian Academy of Sciences, C V Raman Avenue, PB No. 8005, Bangalore 560 080, India.

Guidelines for Authors

Resonance *- journal of science education* is primarily targeted to undergraduate students and teachers. The journal invites contributions in various branches of science and emphasizes a lucid style that will attract readers from diverse backgrounds. A helpful general rule is that at least the first one third of the article should be readily understood by a general audience.

Articles on topics in the undergraduate curriculum, especially those which students often consider difficult to understand, new classroom experiments, emerging techniques and ideas and innovative procedures for teaching specific concepts are particularly welcome. The submitted contributions should not have appeared elsewhere.

Manuscripts should be submitted in *duplicate* to any of the editors. Authors having access to a PC are encouraged to submit an ASCII version on a floppy diskette. If necessary the editors may edit the manuscript substantially in order to maintain uniformity of presentation and to enhance readability. Illustrations and other material if reproduced, must be properly credited; it is the author's responsibility to obtain permission of reproduction (copies of letters of permission should be sent). In case of difficulty, please contact the editors.

Title Authors are encouraged to provide a 4-7 word title and a 4-10 word sub-title. One of these should be a precise technical description of the contents of the article, while the other must attract the general readers' attention.

Author(s) The author's name and mailing address should be provided. A photograph and a brief (in less than 100 words) biographical sketch may be added. Inclusion of phone and fax numbers and e-mail address would help in expediting the processing of manuscripts.

Summary and Brief Provide a 2 to 4 sentence summary, and preferably a one sentence brief for the contents page.

Style and Contents Use simple English. Keep the sentences short. Break up the text into logical units, with readily understandable headings for each. Do not use multiple sub sections. Articles should generally be 1000-2000 words long.

Illustrations Use figures, charts and schemes liberally. A few colour illustrations may be useful. Try to use good quality computer generated images, with neatly labelled axes, clear labels, fonts and shades. Figure captions must be written with care and in some detail. Key features of the illustration may be pointed out in the caption.

Boxes Highlights, summaries, biographical and historical notes and margin notes presented at a level different from the main body of the text and which nevertheless enhance the interest of the main theme can be placed as boxed items. These would be printed in a different typeface. Such a boxed item should fit in a printed page and not exceed 250 words.

Suggested Reading Avoid technical references. If some citations are necessary, mention these as part of the text. A list of suggested readings may be included at the end.

Layout It is preferable to place all the boxes, illustrations and their captions after the main text of the article. The suggested location of the boxes and figures in the printed version may be marked in the text. In the printed version, the main text will occupy two-thirds of each page. The remaining large margin space will be used to highlight the contents of key paragraphs, for figure captions, or perhaps even for small figures. The space is to be used imaginatively to draw attention to the article. Although the editors will attempt to prepare these entries, authors are encouraged to make suitable suggestions and provide them as an annexure.

Book Reviews

The following types of books will be reviewed : (1) text books in subjects of interest to the journal; (2) general books in science brought to the attention of students/teachers; (3) well-known classics; (4) books on educational methods. Books reviewed should generally be affordable to students/teachers (price range Rs.50 to 300).

New books will get preference in review. A list of books received by the academy office will be circulated among the editors who will then decide which ones are to be listed and which to be reviewed.

Sir Karl Raimund Popper was the most outstanding philosopher of science of the century; some even place him along with the great Greek philosophers Plato and Aristotle for the breadth and depth of his knowledge, the capacity to produce original ideas of the utmost importance, and the extraordinary range of application of his ideas, from metaphysics, the philosophy of science, the theory of knowledge to social and political theory.

"I think Popper is incomparably the greatest philosopher of science that has ever been," said Sir Peter Medawar, "...read and meditate upon Popper's writings on the philosophy of science and [...] adopt them as the basis of operation of one's scientific life," was the advice of Sir John Eccles.

ISSN 0971-80

Karl Raimund Popper

(1902-1994)